Proceedings of the

International Congress of Mathematicians

Seoul 2014

SEOUL ICM 2014
INTERNATIONAL CONGRESS OF MATHEMATICIANS

Proceedings of the **International Congress of Mathematicians**

Seoul 2014

VOLUME IV
Invited Lectures

Editors
Sun Young Jang
Young Rock Kim
Dae-Woong Lee
Ikkwon Yie

KM KYUNG MOON SA

Editors

Sun Young Jang, University of Ulsan
Young Rock Kim, Hankuk University of Foreign Studies
Dae-Woong Lee, Chonbuk National University
Ikkwon Yie, Inha University

Technical Editors

Young Rock Kim, The Korean TEX Society
Hyun Woo Kwon, The Korean TEX Society

Proceedings of the International Congress of Mathematicians
August 13–21, 2014, Seoul, Korea

Published by

KYUNG MOON SA Co. Ltd.
174, Wausan-ro Mapo-gu Seoul, Korea
Tel: +82-2-332-2004 Fax: +82-2-336-5193
E-mail: kyungmoon@kyungmoon.com
Homepage: www.kyungmoon.com

ISBN 978-89-6105-807-0
ISBN 978-89-6105-803-2 (set)

Printed in Korea

Contents

14. Mathematical Aspects of Computer Science

15. Numerical Analysis and Scientific Computing

16. Control Theory and Optimizaiton

17. Mathematics in Science and Technology

18. Mathematics Education and Popularization of Mathematics

19. History of Mathematics

12. Probability and Statistics

A short survey of Stein's method

Sourav Chatterjee

Abstract. Stein's method is a powerful technique for proving central limit theorems in probability theory when more straightforward approaches cannot be implemented easily. This article begins with a survey of the historical development of Stein's method and some recent advances. This is followed by a description of a "general purpose" variant of Stein's method that may be called the generalized perturbative approach, and an application of this method to minimal spanning trees. The article concludes with the descriptions of some well known open problems that may possibly be solved by the perturbative approach or some other variant of Stein's method.

Mathematics Subject Classification (2010). Primary 60F05; Secondary 60B10.

Keywords. Stein's method, normal approximation, central limit theorem.

1. Introduction

A sequence of real-valued random variables Z_n is said to converge in distribution to a limiting random variable Z if

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le t) = \mathbb{P}(Z \le t)$$

at all t where the map $t \mapsto \mathbb{P}(Z \le t)$ is continuous. It is equivalent to saying that for all bounded continuous functions g from $\mathbb{R}$ into $\mathbb{R}$ (or into $\mathbb{C}$),

$$\lim_{n\to\infty} \mathbb{E}g(Z_n) = \mathbb{E}g(Z)\,. \tag{1.1}$$

Often, it is not necessary to consider all bounded continuous g, but only g belonging to a smaller class. For example, it suffices to consider all g of the form $g(x) = e^{itx}$, where $i = \sqrt{-1}$ and $t \in \mathbb{R}$ is arbitrary, leading to the method of characteristic functions (that is, Fourier transforms) for proving convergence in distribution.

The case where Z is a normal (alternatively, Gaussian) random variable is of particular interest to probabilists and statisticians, because of the frequency of its appearance as a limit in numerous problems. The normal distribution with mean μ and variance σ is the probability distribution on $\mathbb{R}$ that has probability density

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

with respect to Lebesgue measure. The case $\mu = 0$ and $\sigma = 1$ is called "standard normal" or "standard Gaussian". To show that a sequence of random variables Z_n converges in

Proceedings of the International Congress of Mathematicians, Seoul, 2014

distribution to this Z, one simply has to show that for each t,

$$\lim_{n\to\mathbb{R}} \mathbb{E}(e^{itZ_n}) = \mathbb{E}(e^{itZ}) = e^{it\mu - \sigma^2 t^2/2}.$$

Indeed, this is the most well known approach to proving the classical central limit theorem for sums of independent random variables.

Besides characteristic functions, there are two other classical approaches to proving central limit theorems. First, there is the method of moments, which involves showing that $\lim_{n\to\infty} \mathbb{E}(Z_n^k) = \mathbb{E}(Z^k)$ for every positive integer k. Second, there is an old technique of Lindeberg [54], which has recently regained prominence. I will explain Lindeberg's method in Section 5.

In 1972, Charles Stein [79] proposed a radically different approach to proving convergence to normality. Stein's observation was that the standard normal distribution is the only probability distribution that satisfies the equation

$$\mathbb{E}(Zf(Z)) = \mathbb{E}f'(Z) \tag{1.2}$$

for all absolutely continuous f with a.e. derivative f' such that $\mathbb{E}|f'(Z)| < \infty$. From this, one might expect that if W is a random variable that satisfies the above equation in an approximate sense, then the distribution of W should be close to the standard normal distribution. Stein's approach to making this idea precise was as follows.

Take any bounded measurable function $g : \mathbb{R} \to \mathbb{R}$. Let f be a bounded solution of the differential equation

$$f'(x) - xf(x) = g(x) - \mathbb{E}g(Z), \tag{1.3}$$

where Z is a standard normal random variable. Stein [79] showed that a bounded solution always exists, and therefore for any random variable W,

$$\mathbb{E}g(W) - \mathbb{E}g(Z) = \mathbb{E}(f'(W) - Wf(W)).$$

If the right-hand side is close to zero, so is the left. If we want to consider the supremum of the left-hand side over a class of functions g, then it suffices to do the same on the right for all f obtained from such g. For example, one can prove the following simple proposition:

Proposition 1.1. *Let $\mathcal{D}$ be the set of all $f : \mathbb{R} \to \mathbb{R}$ that are twice continuously differentiable, and $|f(x)| \le 1$, $|f'(x)| \le 1$ and $|f''(x)| \le 1$ for all $x \in \mathbb{R}$. Let Z be a standard normal random variable and W be any random variable. Then*

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(W \le t) - \mathbb{P}(Z \le t)| \le 2\Big(\sup_{f\in\mathcal{D}} |\mathbb{E}(f'(W) - Wf(W))|\Big)^{1/2}.$$

Proof. Fix $\epsilon > 0$. Let $g(x) = 1$ if $x \le t$ and 0 if $x \ge t + \epsilon$, with linear interpolation in the interval $[t, t+\epsilon]$. Let f be a solution of the differential equation (1.3). By standard estimates [36, Lemma 2.4], $|f(x)| \le 2/\epsilon$, $|f'(x)| \le \sqrt{2/\pi}/\epsilon$ and $|f''(x)| \le 2/\epsilon$ for all x. Consequently, $(\epsilon/2)f \in \mathcal{D}$. Since the probability density function of Z is bounded by $1/\sqrt{2\pi}$ everywhere, it follows that

$$\begin{aligned}\mathbb{P}(W \le t) &\le \mathbb{E}g(W)\\ &= \mathbb{E}g(Z) + \mathbb{E}(f'(W) - Wf(W))\end{aligned}$$

$$\leq \mathbb{P}(Z \leq t) + \frac{\epsilon}{\sqrt{2\pi}} + \mathbb{E}(f'(W) - Wf(W))$$
$$\leq \mathbb{P}(Z \leq t) + \frac{\epsilon}{\sqrt{2\pi}} + \frac{2}{\epsilon} \sup_{h \in \mathcal{D}} \mathbb{E}(h'(W) - Wh(W)) \,.$$

Similarly, taking $g(x) = 1$ if $x \leq t - \epsilon$, $g(x) = 0$ if $x \geq t$ and linear interpolation in the interval $[t - \epsilon, t]$, we get

$$\mathbb{P}(W \leq t) \geq \mathbb{P}(Z \leq t) - \frac{\epsilon}{\sqrt{2\pi}} - \frac{2}{\epsilon} \sup_{h \in \mathcal{D}} |\mathbb{E}(h'(W) - Wh(W))| \,.$$

The proof of the proposition is now easily completed by optimizing over ϵ. □

The convenience of dealing with the right-hand side in Proposition 1.1 is that it involves only one random variable, W, instead of the two variables W and Z that occur on the left. This simple yet profound idea gave birth to the field of Stein's method, that has survived the test of time and is still alive as an active field of research within probability theory after forty years of its inception.

2. A brief history of Stein's method

Stein introduced his method of normal approximation in the seminal paper [79] in 1972. The key to Stein's implementation of his idea was the method of exchangeable pairs, devised by Stein in [79]. The key idea is as follows. A pair of random variables or vectors (W, W') is called an *exchangeable pair* if (W, W') has the same distribution as (W', W). Stein's basic idea was that if (W, W') is an exchangeable pair such that for some small number λ,

$$\mathbb{E}(W' - W \mid W) = -\lambda W + o(\lambda) \,,$$
$$\mathbb{E}((W' - W)^2 \mid W) = 2\lambda + o(\lambda) \,, \text{ and}$$
$$\mathbb{E}|W' - W|^3 = o(\lambda) \,,$$

where $o(\lambda)$ denotes random or nonrandom quantities that have typical magnitude much smaller than λ, then X is approximately standard normal. Without going into the precise details, Stein's reasoning goes like this: Given any $f \in \mathcal{D}$ where $\mathcal{D}$ is the function class from Proposition 1.1, it follows by exchangeability that

$$\mathbb{E}((W' - W)(f(W') + f(W))) = 0 \,,$$

because the left-hand side is unchanged if W and W' are exchanged, but it also becomes the negation of itself. But note that by the given conditions,

$$\begin{aligned}\frac{1}{2\lambda}\mathbb{E}((W' - W)(f(W') + f(W))) &= \frac{1}{2\lambda}\mathbb{E}((W' - W)(f(W') - f(W))) \\ &\quad + \frac{1}{\lambda}\mathbb{E}((W' - W)f(W)) \\ &= \frac{1}{2\lambda}\mathbb{E}((W' - W)^2 f'(W)) - \mathbb{E}(Wf(W)) + o(1) \\ &= \mathbb{E}(f'(W)) - \mathbb{E}(Wf(W)) + o(1) \,,\end{aligned}$$

where $o(1)$ denotes a small quantity.

For example, if $W = n^{-1/2}(X_1 + \cdots + X_n)$ for i.i.d. random variables $X_1, \ldots, X_n$ with mean zero, variance one and $\mathbb{E}|X_1|^3 < \infty$, then taking

$$W' = W - \frac{X_I}{\sqrt{n}} + \frac{X_I'}{\sqrt{n}}\,,$$

where I is uniformly chosen from $\{1, \ldots, n\}$ and for each i, X_i' is an independent random variable having the same distribution as X_i, we get an exchangeable pair that satisfies the three criteria listed above with $\lambda = 1/n$ (easy to check).

The monograph [80] also contains the following abstract generalization of the above idea. Suppose that we have two random variables W and Z, and suppose that T_0 is an operator on the space of bounded measurable functions such that $\mathbb{E}T_0 f(Z) = 0$ for all f. Let α be any map that takes a bounded measurable function f on $\mathbb{R}$ to an antisymmetric bounded measurable function αf on $\mathbb{R}^2$ (meaning that $\alpha f(x, y) = -\alpha f(y, x)$ for all x, y).

In the above setting, note that if W' is a random variable such that (W, W') is an exchangeable pair, then $\mathbb{E}\alpha f(W, W') = 0$ for any f. For a function h of two variables, let

$$Th(x) := \mathbb{E}(h(W, W') \mid W = x)\,,$$

so that $\mathbb{E}T\alpha f(W) = \mathbb{E}\alpha f(W, W') = 0$ for any f. Consequently, given g, if f is a solution of the functional equation

$$T_0 f(x) = g(x) - \mathbb{E}g(Z)\,,$$

then

$$\mathbb{E}g(W) - \mathbb{E}g(Z) = \mathbb{E}T_0 f(W) = \mathbb{E}(T_0 f(W) - T\alpha f(W))\,. \tag{2.1}$$

Thus, if $T_0 \approx T\alpha$, then Z and W have approximately the same distributions. For example, for normal approximation, we can take

$$T_0 f(x) = f'(x) - xf(x) \text{ and } \alpha f(x, y) = (2\lambda)^{-1}(x - y)(f(x) + f(y))$$

, where λ is as above. If the three conditions listed by Stein hold for an exchangeable pair (W, W'), then indeed $T_0 \approx T\alpha$, as we have shown above.

The identity (2.1) is the content of a famous commutative diagram of Stein [80]. It has been used in contexts other than normal approximation — for example, for Poisson approximation in [26] and for the analysis of Markov chains in [39].

A notable success story of Stein's method was authored by Bolthausen [13] in 1984, when he used a sophisticated version of the method of exchangeable pairs to obtain an error bound in a famous combinatorial central limit theorem of Hoeffding. The problem here is to prove a central limit theorem for an object like $W = \sum_{i=1}^n a_{i\pi(i)}$, where a_{ij} is a given array of real numbers, and π is a uniform random permutation of $\{1, \ldots, n\}$. Bolthausen defined

$$W' = W - a_{I\pi(I)} - a_{J\pi(J)} + a_{I\pi(J)} + a_{J\pi(I)}\,,$$

and proved that (W, W') is an exchangeable pair satisfying the three required conditions. The difficult part in Bolthausen's work was to derive a sharp error bound, since the error rate given by a result like Proposition 1.1 is usually not optimal.

Incidentally, it has been proved recently by Röllin [74] that to apply exchangeable pairs for normal approximation, it is actually not necessary that W and W' are exchangeable; one can make an argument go through if W and W' have the same distribution.

Stein's 1986 monograph [80] was the first book-length treatment of Stein's method. After the publication of [80], the field was given a boost by the popularization of the method of dependency graphs by Baldi and Rinott [6], a striking application to the number of local maxima of random functions by Baldi, Rinott and Stein [7], and central limit theorems for random graphs by Barbour, Karoński and Ruciński [11], all in 1989.

The method of dependency graphs, as a version of Stein's method, was introduced in Louis Chen's 1971 Ph.D. thesis on Poisson approximation and the subsequent publication [32]. It was developed further by Chen [33] before being brought to wider attention by Baldi and Rinott [6]. Briefly, the method may be described as follows. Suppose that $(X_i)_{i\in V}$ is a collection of random variables indexed by some finite set V. A *dependency graph* is an undirected graph on the vertex set V such that if A and B are two subsets of V such that there are no edges with one endpoint in A and the other in B, then the collections $(X_i)_{i\in A}$ and $(X_i)_{i\in B}$ are independent. Fix a dependency graph, and for each i, let N_i be the neighborhood of i in this graph, including the vertex i. Let $W = \sum_{i\in V} X_i$ and assume that $\mathbb{E}(X_i) = 0$ for each i. Define

$$W_i := \sum_{j\notin N_i} X_j\,,$$

so that W_i is independent of X_i. Then note that for any smooth f,

$$\begin{aligned}\mathbb{E}(Wf(W)) &= \sum_{i\in V}\mathbb{E}(X_i f(W))\\ &= \sum_{i\in V}\mathbb{E}(X_i(f(W)-f(W_i)))\\ &\approx \sum_{i\in V}\mathbb{E}(X_i(W-W_i)f'(W)) = \mathbb{E}\biggl(\Bigl(\sum_{i\in V}X_i(W-W_i)\Bigr)f'(W)\biggr)\,,\end{aligned}$$

where the approximation holds under the condition that $W \approx W_i$ for each i. Define $T := \sum_{i\in V} X_i(W - W_i)$. Let $\sigma^2 := \mathbb{E}T$. The above approximation, when valid, implies that $\mathrm{Var}W = \mathbb{E}W^2 \approx \sigma^2$. Therefore if T has a small variance, then $\mathbb{E}(Wf(W)) \approx \sigma^2\mathbb{E}f'(W)$. By a slight variant of Proposition 1.1, this shows that W is approximately normal with mean zero and variance σ^2.

To gain a hands-on understanding of the dependency graph method, the reader can check that this technique works when $Y_1, \ldots, Y_n$ are independent random variables with mean zero, and $X_i = n^{-1/2}Y_iY_{i+1}$ for $i = 1, \ldots, n-1$. Here $V = \{1, \ldots, n-1\}$, and a dependency graph may be defined by putting an edge between i and j whenever $|i-j| = 1$.

The new surge of activity that began in the late eighties continued through the nineties, with important contributions coming from Barbour [8] in 1990, who introduced the diffusion approach to Stein's method; Avram and Bertsimas [5] in 1993, who applied Stein's method to solve an array of important problems in geometric probability; Goldstein and Rinott [50] in 1996, who developed the method of size-biased couplings for Stein's method, improving on earlier insights of Baldi, Rinott and Stein [7]; Goldstein and Reinert [49] in 1997, who introduced the method of zero-bias couplings; and Rinott and Rotar [72] in 1997, who solved a well known open problem related to the antivoter model using Stein's method. Sometime later, in 2004, Chen and Shao [38] did an in-depth study of the dependency graph approach, producing optimal Berry-Esséen type error bounds in a wide range of problems. The 2003

monograph of Penrose [66] gave extensive applications of the dependency graph approach to problems in geometric probability.

I will now try to outline the basic concepts behind some of the methods cited in the preceding paragraph.

The central idea behind Barbour's diffusion approach [8] is that if a probability measure μ on some abstract space is the unique invariant measure for a diffusion process with generator $\mathcal{L}$, then under mild conditions μ is the only probability measure satisfying $\int \mathcal{L}f d\mu = 0$ for all f in the domain of $\mathcal{L}$; therefore, if a probability measure ν has the property that $\int \mathcal{L}f d\nu \approx 0$ in some suitable sense for a large class of f's, then one may expect that ν is close to μ is some appropriate metric. Generalizing Stein's original approach, Barbour then proposed the following route to make this idea precise. Given a function g on this abstract space, one can try to solve for

$$\mathcal{L}f(x) = g(x) - \int g d\mu\,,$$

and use

$$\int g d\nu - \int g d\mu = \int \mathcal{L}f d\nu \approx 0\,.$$

To see how Stein's method of normal approximation fits into this picture, one needs to recall that the standard normal distribution on $\mathbb{R}$ is the unique invariant measure for a diffusion process known as the *Ornstein-Uhlenbeck process*, whose generator is $\mathcal{L}f(x) = f''(x) - xf'(x)$. This looks different than the original Stein operator $f'(x) - xf(x)$, but it is essentially the same: one has to simply replace f by f' and f' by f''.

In [8], Barbour used this variant of Stein's method to solve some problems about diffusion approximation. However, the most significant contribution of Barbour's paper was a clarification of the mysterious nature of the method of exchangeable pairs. A one dimensional diffusion process $(X_t)_{t\geq 0}$ with drift coefficient $a(x)$ and diffusion coefficient $b(x)$ is a continuous time stochastic process adapted to some filtration $\{\mathcal{F}_t\}_{t\geq 0}$ satisfying, as $h \to 0$,

$$\begin{aligned}
&\mathbb{E}(X_{t+h} - X_t \mid \mathcal{F}_t) = a(X_t)h + o(h)\,,\\
&\mathbb{E}((X_{t+h} - X_t)^2 \mid \mathcal{F}_t) = b(X_t)^2 h + o(h)\,, \text{ and}\\
&\mathbb{E}|X_{t+h} - X_t|^3 = o(h)\,.
\end{aligned}$$

An exchangeable pair (W, W') naturally defines a stationary, reversible Markov chain W_0, $W_1, W_2, \ldots$, where $W_0 = W$, $W_1 = W'$, and for each i, the conditional distribution of W_{i+1} given W_i is the same as that of W_1 given W_0. If the pair (W, W') satisfies the three conditions listed by Stein for some small λ, then in a scaling limit as $\lambda \to 0$, the Markov chain defined above converges to a diffusion process with drift function $a(x) = -x$ and diffusion coefficient $\sqrt{2}$. This is precisely the standard Ornstein-Uhlenbeck process whose stationary distribution is the standard normal. Therefore one can expect that W is approximately normally distributed. Note that this argument is quite general, and not restricted to normal approximation. In a later paragraph, I will briefly point out some generalizations of Stein's method using Barbour's approach.

The method of size-biased couplings in Stein's method was introduced in the paper of Baldi, Rinott and Stein [7], and was fully developed by Goldstein and Rinott [50]. The size-biased transform of a non-negative random variable W with mean λ is a random variable,

usually denoted by W^*, such that for all g,

$$\mathbb{E}(Wg(W)) = \lambda \mathbb{E}g(W^*)\,.$$

Size biasing is actually a map on probability measures, which takes a probability measure μ on the non-negative reals to a probability measure ν defined as $d\nu(x) = \lambda^{-1} x d\mu(x)$, where λ is the mean of μ. Size biasing is an old concept, predating Stein's method, probably originating in the survey sampling literature. (Actually, the name "size-biasing" comes from the survey sampling procedure where a sample point is chosen with probability proportional to some notion of size.) As a consequence of its classical origins and usefulness in a variety of domains, there are many standard procedures to construct size-biased versions of complicated random variables starting from simpler ones. For example, if $X_1, \ldots, X_n$ are i.i.d. non-negative random variables, and $W = X_1 + \cdots + X_n$, and X_1^* is a size-biased version of X_1, then $W^* = X_1^* + X_2 + \cdots + X_n$ is a size-biased version of W. To see this, just note that for any g,

$$\begin{aligned}\mathbb{E}(Wg(W)) &= n\mathbb{E}(X_1 g(X_1 + \cdots X_n)) \\ &= n\mathbb{E}(X_1)\mathbb{E}g(X_1^* + X_2 + \cdots + X_n) \\ &= \mathbb{E}(W)\mathbb{E}g(W^*)\,.\end{aligned}$$

For more complicated examples, see [50].

In Stein's method, size biasing is used in the following manner: Suppose that W is a non-negative random variable with mean λ and variance σ^2. Suppose that we are able to construct a size-biased version W^* of W on the same probability space, such that

$$\begin{aligned}\mathbb{E}(W^* - W \mid W) &= \frac{\sigma^2}{\lambda}(1 + o(1))\,, \text{ and} \\ \mathbb{E}(W^* - W)^2 &= o\left(\frac{\sigma^3}{\lambda}\right).\end{aligned}$$

Then the standardized random variable $X := (W - \lambda)/\sigma$ is approximately standard normal. To understand why this works, let $Y := (W^* - \lambda)/\sigma$ and note that under the two conditions displayed above,

$$\begin{aligned}\mathbb{E}(Xf(X)) &= \frac{1}{\sigma}\mathbb{E}(Wf(X)) - \frac{\lambda}{\sigma}\mathbb{E}f(X) \\ &= \frac{\lambda}{\sigma}\mathbb{E}(f(Y) - f(X)) \\ &= \frac{\lambda}{\sigma}\mathbb{E}((Y - X)f'(X)) + \frac{\lambda}{\sigma}O(\mathbb{E}(Y - X)^2) \\ &= \frac{\lambda}{\sigma^2}\mathbb{E}(\mathbb{E}(W^* - W|W)f'(X)) + \frac{\lambda}{\sigma^3}O(\mathbb{E}(W^* - W)^2) \\ &= \mathbb{E}f'(X) + o(1)\,.\end{aligned}$$

For a mathematically precise version of the above argument, see [50, Theorem 1.1].

The method of size biased couplings is quite a powerful tool for proving central limit theorems for non-negative random variables, especially those that arise as sums of mildly dependent variables. The only hurdle is that one has to be able to construct a suitable size-biased

coupling. There is also the other limitation that W has to be non-negative. To overcome these limitations, Goldstein and Reinert [49] introduced the method of zero-bias couplings. Given a random variable W with mean zero and variance σ^2, the zero-biased transform W' of W is a random variable satisfying

$$\mathbb{E}(Wf(W)) = \sigma^2 \mathbb{E} f'(W')$$

for all differentiable f whenever the left-hand side is well-defined. It is clear from Proposition 1.1 that if one can define a zero-bias transform W' on the same probability space as W such that $W' \approx W$ with high probability, then W is approximately normal with mean 0 and variance σ^2. The construction of zero-bias transforms can be quite tricky. The method has been systematically developed and used to solve a variety of problems by a number of authors, starting with Goldstein and Reinert [49].

A feature of Stein's method of normal approximation that has limited its applicability throughout the history of the subject is that it works only for problems where "something nice" happens. This is true of all classical versions of the method, such as the method of exchangeable pairs, the dependency graph approach, size-biased couplings and zero-bias couplings. For exchangeable pairs, we need that the three conditions listed by Stein are valid. For dependency graphs, we need the presence of a dependency graph of relatively small degree. For the coupling techniques, we need to be able to construct the couplings. Given a general problem with no special structure, it is often difficult to make these methods work. Intending to come up with a more general approach, I introduced a new method in 2008 in the paper [21] for discrete systems, and a corresponding continuous version in [22] in 2009. This new approach (which I am calling *the generalized perturbative approach* in this article) was used to solve a number of questions in geometric probability in [21], random matrix central limit theorems in [22], number theoretic central limit theorems in [31], and an error bound in a central limit theorem for minimal spanning trees in [29]. The generalized perturbative method is described in detail in Section 3.

The paper [22] also introduced the notion of *second order Poincaré inequalities*. The simplest second order Poincaré inequality, derived in [22], states that if $X = (X_1, \ldots, X_n)$ is a vector of i.i.d. standard normal random variables, $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function with gradient ∇f and Hessian matrix $\mathrm{Hess} f$, and $W := f(X)$ has mean zero and variance 1, then

$$\sup_{A \in \mathcal{B}(\mathbb{R})} |\mathbb{P}(W \in A) - \mathbb{P}(Z \in A)| \le 2\sqrt{5} (\mathbb{E}\|\nabla f(X)\|^4)^{1/4} (\mathbb{E}\|\mathrm{Hess} f(X)\|_{\mathrm{op}}^4)^{1/4},$$

where $\|\nabla f(X)\|$ is the Euclidean norm of $\nabla f(X)$, $\|\mathrm{Hess} f(X)\|_{\mathrm{op}}$ is the operator norm of $\mathrm{Hess} f(X)$, and $\mathcal{B}(\mathbb{R})$ is the set of Borel subsets of $\mathbb{R}$. In [22], this inequality was used to prove new central limit theorems for linear statistics of eigenvalues of random matrices. The name "second order Poincaré inequality" is inspired from the analogy with the usual Poincaré inequality for the normal distribution, which states that $\mathrm{Var} f(X) \le \mathbb{E}\|\nabla f(X)\|^2$ for any absolutely continuous f. Although this does not look like anything related to Stein's method, a close inspection of the proof in [22] makes it clear that it is in fact an offshoot of Stein's method.

Incidentally, the usual Poincaré inequality has also been used to prove central limi theorems, for example by Chen [34], using a characterization of the normal distribution by Borovkov and Utev [15].

Second order Poincaré inequalities have been useful in several subsequent works, e.g. in Nourdin, Peccati and Reinert [62], Nolen [59], etc. Indeed, it may be said that the whole thriving area of Stein's method in Malliavin calculus, pioneered by Nourdin and Peccati [60], is an "abstractification" of the ideas contained in [21] and [22]. The new method was later unified with other branches of Stein's method through the concept of *Stein couplings* introduced by Chen and Röllin [37].

Normal approximation is not the only area covered by Stein's method. In 1975, Louis Chen [32] devised a version of Stein's method for Poisson approximation, expanding on his 1971 Ph.D. thesis under Stein. The *Chen-Stein method* of Poisson approximation is a very useful tool in its own right, finding applications in many areas of the applied sciences. The main idea is that a Poisson random variable X with mean λ is the only kind of random variable satisfying

$$\mathbb{E}(Xf(X)) = \lambda \mathbb{E}f(X+1)$$

for every f, and then proceed from there as usual by developing a suitable version of Proposition 1.1. The subject of Poisson approximation by Stein's method took off with the papers of Arratia, Goldstein and Gordon [3, 4] and the classic text of Barbour, Holst and Janson [10], all appearing in the period between 1989 and 1992. A relatively recent survey of Poisson approximation by Stein's method is given in my paper [26] with Diaconis and Meckes.

Besides normal and Poisson, Stein's method has been used sometimes for other kinds of distributional approximations. One basic idea was already available in Stein's 1986 monograph [80], and a different one in Barbour's paper [8] on the diffusion approach to Stein's method. These ideas were implemented in various forms by Mann [57] in 1994 for chi-square approximation, Luk [55] in 1997 for gamma approximation, Holmes [52] in 2004 for birth-and-death chains, and Reinert [68] in 2005 for approximation of general densities. In 2005, Fulman [46] extended the method of exchangeable pairs to study Plancherel measures on symmetric groups. Stein's method for a mixture of two normal distributions, with an application to spin glasses, appeared in my 2010 paper [23], while another non-normal distribution arising at the critical temperature of the Curie-Weiss model of ferromagnets was tackled in my joint paper with Shao [30] in 2011 and in a paper of Eichelsbacher and Löwe [41] in 2010. Several papers on Stein's method for geometric and exponential approximations have appeared in the literature, including an early paper of Peköz [63] from 1996, a paper of myself with Fulman and Röllin [27] that appeared in 2011, and papers of Peköz and Röllin [64] and Peköz, Röllin and Ross [65] that appeared in 2011 and 2013 respectively.

Another area of active research is Stein's method for multivariate normal approximation. Successful implementations were carried out by Götze [51] in 1991, Bolthausen and Götze [14] in 1993, and Rinott and Rotar [71] in 1996. The complexities of Götze's method were clarified by Bhattacharya and Holmes [12] in 2010. In a joint paper [28] with Meckes in 2008, we found a way to implement the method of exchangeable pairs in the multivariate setting. The main idea here is to generalize Barbour's diffusion approach to the multidimensional setting, by considering the multivariate Ornstein-Uhlenbeck process and the related semigroup. This naturally suggests a multivariate generalization of the three exchangeable pair conditions listed by Stein. The relevant generalization of the Stein equation (1.3), therefore, is

$$\Delta f(x) - x \cdot \nabla f(x) = g(x) - \mathbb{E}g(Z)\,,$$

where Δf is the Laplacian of f, ∇f is the gradient of f, $x \cdot \nabla f(x)$ is the inner product of the vector x and the gradient vector $\nabla f(x)$, and Z is a multidimensional standard normal

random vector. The method was greatly advanced, with many applications, by Reinert and Röllin [69, 70] in 2009 and 2010. Further advances were made in the recent manuscript of Röllin [75].

Incidentally, there is a rich classical area of multivariate normal approximation, and a lot of energy spent on what class of sets the approximation holds for. This remains to be worked out for Stein's method.

Besides distributional approximations, Stein's method has also been used to prove concentration inequalities. Preliminary attempts towards deviation inequalities were made by Stein in his 1986 monograph [80], which were somewhat taken forward by Raič in 2007. The first widely applicable set of concentration inequalities using Stein's method of exchangeable pairs appeared in my Ph.D. thesis [18] in 2005, some of which were collected together in the 2007 paper [20]. A more complex set of examples was worked out in a later paper with Dey [25] in 2010. One of the main results of [18, 20] is that if (W, W') is an exchangeable pair of random variables and $F(W, W')$ is an antisymmetric function of (W, W') (meaning that $F(W, W') = -F(W', W)$), then for all $t \geq 0$,

$$\mathbb{P}(|f(W)| \geq t) \leq 2e^{-t^2/2C},$$

where $f(W) = \mathbb{E}(F(W, W')|W)$ and C is a number such that

$$|(f(W) - f(W'))F(W, W')| \leq C \text{ with probability one.}$$

Surprisingly, this abstract machinery has found quite a bit of use in real applications. In 2012, Mackey and coauthors [56] extended the method to the domain of matrix concentration inequalities, thereby solving some problems in theoretical machine learning. In 2011, Ghosh and Goldstein [47, 48] figured out a way to use size-biased couplings for concentration inequalities.

There are a number of nonstandard applications of Stein's method that have not yet gathered a lot of follow up action, for example, Edgeworth expansions (Rinott and Rotar [73]), rates of convergence of Markov chains (Diaconis [39]), strong approximation in the style of the KMT embedding theorem (my paper [24]), moderate deviations (Chen et al. [35]) and even in the analysis of simulations (Stein et al. [81]). A great deal of hard work has gone into proving sharp Berry-Esséen bounds using Stein's method. Some of this literature is surveyed in Chen and Shao [38].

A number of well written monographs dedicated to various aspects of Stein's method are in existence. The book of Barbour, Holst and Janson [10] is a classic text on Poisson approximation by Stein's method. The recent monograph by Chen, Goldstein and Shao [36] is a very readable and comprehensive account of normal approximation by Stein's method. The survey of Ross [76], covering many aspects of Stein's method, is already attaining the status of a must-read in this area. The monograph [61] of Nourdin and Peccati describes the applications of Stein's method in Malliavin calculus. The edited volumes [9] and [40] are also worth a look.

Lastly, I should clarify that the above review was an attempt to cover only the theoretical advances in Stein's method. The method has found many applications in statistics, engineering, machine learning, and other areas of applications of mathematics. I have made no attempt to survey these applications.

This concludes my very rapid survey of existing techniques and ideas in Stein's method. I apologize to anyone whose work I may have inadvertently left out. In the rest of this

manuscript, I will attempt to briefly explain the generalized perturbative method introduced in the papers [21] and [22], and then conclude by stating some open problems.

3. The generalized perturbative approach

Let $\mathcal{X}$ be a measure space and suppose $X = (X_1, \dots, X_n)$ is a vector of independent $\mathcal{X}$-valued random variables. Let $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function and let $W := f(X)$. Suppose that $\mathbb{E}W = 0$ and $\mathbb{E}W^2 = 1$. I will now outline a general technique for getting an upper bound on the distance of W from the standard normal distribution using information about how f changes when one coordinate of X is perturbed. Such techniques have long been commonplace in the field of concentration inequalities. Suitable versions were introduced for the first time in the context of normal approximation in the papers [21, 22]. I am now calling this the *generalized perturbative approach* to Stein's method. The word "generalized" is added to the name because the method of exchangeable pairs is also a perturbative approach, but this is more general.

Let $X' = (X'_1, \dots, X'_n)$ be an independent copy of X. Let $[n] = \{1, \dots, n\}$, and for each $A \subseteq [n]$, define the random vector X^A as

$$X_i^A = \begin{cases} X'_i & \text{if } i \in A, \\ X_i & \text{if } i \notin A. \end{cases}$$

When A is singleton set like $\{i\}$, write X^i instead of $X^{\{i\}}$. Similarly, write $A \cup i$ instead of $A \cup \{i\}$. Define a randomized derivative of f along the ith coordinate as

$$\Delta_i f := f(X) - f(X^i)\,,$$

and for each $A \subseteq [n]$ and $i \notin A$, let

$$\Delta_i f^A := f(X^A) - f(X^{A \cup i})\,.$$

For each proper subset A of $[n]$ define

$$\nu(A) := \frac{1}{n\binom{n-1}{|A|}}\,.$$

Note that when restricted to the set of all subsets of $[n]\backslash\{i\}$ for some given i, ν is a probability measure. Define

$$T := \frac{1}{2}\sum_{i=1}^{n} \sum_{A \subseteq [n]\backslash\{i\}} \nu(A)\Delta_i f \Delta_i f^A\,.$$

The generalized perturbative approach is based on the following completely general upper bound on the distance of W from normality using the properties of the discrete derivatives $\Delta_i f$ and $\Delta_i f^A$.

Theorem 3.1 (Variant of Theorem 2.2 in [21]). *Let W be as above and Z be a standard normal random variable. Then*

$$\sup_{t \in \mathbb{R}} |\mathbb{P}(W \le t) - \mathbb{P}(Z \le t)| \le 2\left(\sqrt{\mathrm{Var}(\mathbb{E}(T|W))} + \frac{1}{4}\sum_{i=1}^{n} \mathbb{E}|\Delta_i f|^3\right)^{1/2}.$$

In practice, the variance of $\mathbb{E}(T|W)$ may be upper bounded by the variance of $\mathbb{E}(T|X)$ or the variance of T, which are easier to handle mathematically.

The following simple corollary may often be useful for problems with local dependence. We will see an application of this to minimal spanning trees in Section 4.

Corollary 3.2. *Consider the setting of Theorem 3.1. For each i, j, let c_{ij} be a constant such that for all $A \subseteq [n]\backslash\{i\}$ and $B \subseteq [n]\backslash\{j\}$,*

$$\mathrm{Cov}(\Delta_i f \Delta_i f^A,\, \Delta_j f \Delta_j f^B) \le c_{ij}\,.$$

Then

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(W \le t) - \mathbb{P}(Z \le t)| \le \sqrt{2}\biggl(\sum_{i,j=1}^n c_{ij}\biggr)^{1/4} + \biggl(\sum_{i=1}^n \mathbb{E}|\Delta_i f|^3\biggr)^{1/2}.$$

Intuitively, the above corollary says that if most pairs of discrete derivatives are approximately independent, then W is approximately normal. This condition may be called *the approximate independence of small perturbations.*

For example, if $X_1, \ldots, X_n$ are real-valued with mean zero and variance one, and $W = n^{-1/2}\sum X_i$, then we may take $c_{ij} = 0$ when $i \ne j$ and $c_{ii} = C/n^2$ for some constant C depending on the distribution of the X_i's. Moreover note that $|\Delta_i f|$ is of order $n^{-1/2}$. Therefore, Corollary 3.2 gives a proof of the ordinary central limit theorem for sums of i.i.d. random variables with an $n^{-1/4}$ rate of convergence. This rate is suboptimal, but this suboptimality is a general feature Stein's method, requiring quite a bit of effort to overcome.

Theorem 3.1 was used to solve several questions in geometric probability (related to nearest neighbor distances and applications in statistics) in [21], prove a number theoretic central limit theorem in [31] and obtain a rate of convergence in a central limit theorem for minimal spanning trees in [29]. When $X_1, \ldots, X_n$ are i.i.d. normal random variables, a "continuous" version of this theorem, where the perturbations are done in a continuous manner instead of replacing by independent copies, was proved in [22]. This continuous version of Theorem 3.1 was then used to derive the so-called second order Poincaré inequality for the Gaussian distribution.

The remainder of this section is devoted to the proofs of Theorem 3.1 and Corollary 3.2. Applications are worked out in the subsequent sections.

Proof of Theorem 3.1. Consider the sum

$$\sum_{i=1}^n \sum_{A\subseteq[n]\backslash\{i\}} \nu(A)\Delta_i f^A\,.$$

Clearly, this is a linear combination of $\{f(X^A), A \subseteq [n]\}$. It is a matter of simple verification that the positive and negative coefficients of $f(X^A)$ in this linear combination cancel out except when $A = [n]$ or $A = \emptyset$. In fact, the above expression is identically equal to $f(X) - f(X')$.

Let $g : \mathcal{X} \to \mathbb{R}$ be another measurable function. Fix A and $i \notin A$, and let $U = g(X)\Delta_i f^A$. Then U is a function of the random vectors X and X'. The joint distribution of (X, X') remains unchanged if we interchange X_i and X_i'. Under this operation, U changes to $U' := -g(X^i)\Delta_i f^A$. Thus,

$$\mathbb{E}(U) = \mathbb{E}(U') = \frac{1}{2}\mathbb{E}(U + U') = \frac{1}{2}\mathbb{E}\bigl(\Delta_i g \Delta_i f^A\bigr)\,.$$

As a consequence of the above steps and the assumption that $\mathbb{E}W = 0$, we arrive at the identity

$$\begin{aligned}\mathbb{E}(g(X)W) &= \mathbb{E}\big(g(X)(f(X) - f(X'))\big) \\ &= \mathbb{E}\Big(\sum_{i=1}^n \sum_{A \subseteq [n]\setminus\{i\}} \nu(A) g(X) \Delta_i f\Big) \\ &= \frac{1}{2}\mathbb{E}\Big(\sum_{i=1}^n \sum_{A \subseteq [n]\setminus\{i\}} \nu(A) \Delta_i g \Delta_i f^A\Big)\,.\end{aligned}$$

In particular, taking $g = f$ gives $\mathbb{E}T = \mathbb{E}W^2 = 1$. Next, take any $\varphi : \mathbb{R} \to \mathbb{R}$ that belongs to the class $\mathcal{D}$ defined in Proposition 1.1. Let $g := \varphi \circ f$. By the above identity,

$$\mathbb{E}(\varphi(W)W) = \frac{1}{2}\sum_{i=1}^n \sum_{A \subseteq [n]\setminus\{i\}} \nu(A)\mathbb{E}(\Delta_i g \Delta_i f^A).$$

By the mean value theorem and the fact that $|\varphi''(x)| \le 1$ for all x,

$$\mathbb{E}|\Delta_i g \Delta_i f^A - \varphi'(W)\Delta_i f \Delta_i f^A| \le \frac{1}{2}\mathbb{E}|(\Delta_i f)^2 \Delta_i f^A| \le \frac{1}{2}\mathbb{E}|\Delta_i f|^3\,,$$

where the last step follows by Hölder's inequality. Combining the last two displays gives

$$|\mathbb{E}(\varphi(W)W) - \mathbb{E}(\varphi'(W)T)| \le \frac{1}{4}\sum_{i=1}^n \sum_{A \subseteq [n]\setminus\{i\}} \nu(A)\mathbb{E}|\Delta_i f|^3 = \frac{1}{4}\sum_{i=1}^n \mathbb{E}|\Delta_i f|^3\,.$$

Next, note that since $\mathbb{E}T = 1$ and $|\varphi'(x)| \le 1$ for all x,

$$\begin{aligned}|\mathbb{E}(\varphi'(W)T) - \mathbb{E}\varphi'(W)| &= |\mathbb{E}(\varphi'(W)(\mathbb{E}(T|W) - 1))| \\ &\le \mathbb{E}|\mathbb{E}(T|W) - 1| \le \sqrt{\mathrm{Var}(\mathbb{E}(T|W))}\,.\end{aligned}$$

By the last two displays,

$$|\mathbb{E}(\varphi(W)W - \varphi'(W))| \le \sqrt{\mathrm{Var}(\mathbb{E}(T|W))} + \frac{1}{4}\sum_{i=1}^n \mathbb{E}|\Delta_i f|^3\,.$$

Since this is true for any $\varphi \in \mathcal{D}$, Proposition 1.1 completes the proof of Theorem 3.1. □

Proof of Corollary 3.2. Observe that

$$\begin{aligned}\mathrm{Var}T &\le \frac{1}{4}\sum_{i,j=1}^n \sum_{\substack{A \subseteq [n]\setminus\{i\} \\ B \subseteq [n]\setminus\{j\}}} \nu(A)\nu(B)\,\mathrm{Cov}(\Delta_i f \Delta_i f^A,\, \Delta_j f \Delta_j f^B) \\ &\le \frac{1}{4}\sum_{i,j=1}^n \sum_{\substack{A \subseteq [n]\setminus\{i\} \\ B \subseteq [n]\setminus\{j\}}} \nu(A)\nu(B)\,c_{ij} = \frac{1}{4}\sum_{i,j=1}^n c_{ij}\,.\end{aligned}$$

To complete the proof, apply Theorem 3.1 and the inequality $(x+y)^{1/2} \le x^{1/2} + y^{1/2}$ to separate out the two terms in the error bound. □

4. Application to minimal spanning trees

In this section, I will describe an application of the generalized perturbative method to prove a central limit theorem for minimal spanning trees on lattices with random edge weights. This is a small subset of a joint work with Sen [29]. The major objective of [29] was to obtain a rate of convergence, using the generalized perturbative approach, in a central limit theorem for the Euclidean minimal spanning tree due to Kesten and Lee [53]. Kesten and Lee used the martingale central limit theorem to solve this problem (without an error bound), which was a long-standing open question at the time of its solution (except for the two-dimensional case, which was solved by Alexander [2]). My interest in this area stemmed from a quest to understand normal approximation in random combinatorial optimization. Many such problems are still wide open. I will talk about some of them in the next section.

Let E be the set of edges of the integer lattice $\mathbb{Z}^d$. Let $(\omega_e)_{e\in E}$ be a set of i.i.d. edge weights, drawn from a continuous probability distribution on the positive real numbers with bounded support. For each n, let V_n be the set $[-n,n]^d \cap \mathbb{Z}^d$, and let E_n be the set of edges of V_n. The *minimal spanning tree* on the graph $G_n = (V_n, E_n)$ with edge weights $(\omega_e)_{e\in E_n}$ is the spanning tree that minimizes the sum of edge weights. Since the edge-weight distribution is continuous, this tree is unique.

Let M_n be the sum of edge weights of the minimal spanning tree on G_n. We will now see how to use Corollary 3.2 to give a simple proof of the following central limit theorem for M_n.

Theorem 4.1 (Corollary of Theorem 2.4 in [29]). *Let $\mu_n := \mathbb{E}M_n$, $\sigma_n^2 := \mathrm{Var}M_n$, and*

$$f_n = f_n((\omega_e)_{e\in E_n}) := \frac{M_n - \mu_n}{\sigma_n},$$

so that f_n is a standardized version of M_n, with mean zero and variance one. Then f_n converges in law to the standard normal distribution as n goes to infinity.

Note that the above theorem does not have a rate of convergence. Theorem 2.4 in [29] has an explicit rate of convergence, but the derivation of that rate will take us too far afield; moreover that will be an unnecessary digression from the main purpose of this section, which is to demonstrate a nontrivial application of the generalized perturbative approach. In the remainder of this section, I will present a short proof of Theorem 4.1 using the version of the generalized perturbative approach given in Corollary 3.2.

To apply Corollary 3.2, we first have to understand how M_n changes when one edge weight is replaced by an independent copy. This is a purely combinatorial issue. Following the notation of the previous section, I will denote the difference by $\Delta_e M_n$. The goal, eventually, is to show that $\Delta_e M_n$ is approximately equal to a quantity that depends only on some kind of a local neighborhood of e. This will allow us to conclude that the covariances in Corollary 3.2 are small. The following lemma gives a useful formula for the discrete derivative $\Delta_e M_n$, which is a first step towards this eventual goal.

Lemma 4.2. *For each edge $e \in E$ and each n such that $e \in E_n$, let $\alpha_{e,n}$ denote the smallest real number α such that there is a path from one endpoint of e to the other, lying entirely in V_n but not containing the edge e, such that all edges on this path have weight $\le \alpha$. If the edge weight ω_e is replaced by an independent copy ω_e', and $\Delta_e M_n$ denotes the resulting change in M_n, then $\Delta_e M_n = (\alpha_{e,n} - \omega_e')^+ - (\alpha_{e,n} - \omega_e)^+$ where x^+ denotes the positive part of x.*

To prove this lemma, we first need to prove a well known characterization of the minimal spanning tree on a graph with distinct edge weights. Since we have assumed that the edge weight distribution is continuous, the weights of all edges and paths are automatically distinct with probability one.

Lemma 4.3. *An edge $e \in E_n$ belongs to the minimal spanning tree on G_n if and only if $\omega_e < \alpha_{e,n}$. Moreover, if h is the unique edge with weight $\alpha_{e,n}$, then the lighter of the two edges e and h belongs to the tree and the other one does not.*

Proof. Let T denote the minimal spanning tree. First suppose that $e \in T$. Let T_1 and T_2 denote the two connected components of $T\backslash\{e\}$. There is a path in G_n connecting the two endpoints of e, which does not contain e and whose edge weights are all $\leq \alpha_{e,n}$. At least one edge r in this path is a bridge from T_1 to T_2. If $\omega_e > \alpha_{e,n}$, then we can delete the edge e from T and add the edge r to get a tree that has total weight $< M_n$, which is impossible. Therefore $\omega_e < \alpha_{e,n}$. Next, suppose that $\omega_e < \alpha_{e,n}$. Let P be the unique path in T that connects the two endpoints of e. If P does not contain e, then P must contain an edge that has weight $\geq \alpha_{e,n} > \omega_e$. Deleting this edge from T and adding the edge e gives a tree with weight $< M_n$, which is impossible. Hence T must contain e.

To prove the second assertion of the lemma, first observe that if $\omega_h > \omega_e$, then $e \in T$ and $h \notin T$ by the first part. On the other hand if $\omega_h < \omega_e$, then $e \notin T$ by the first part; and if $\alpha_{h,n} < \omega_h$, then there exists a path connecting the two endpoints of e whose edge weights are all $< \alpha_{e,n}$, which is impossible. Therefore again by the first part, $h \in T$. □

We are now ready to prove Lemma 4.2.

Proof of Lemma 4.2. Let T and T' denote the minimal spanning trees before and after replacing ω_e by ω'_e. Note that since T and T' are both spanning trees, we have (I): T and T' must necessarily have the same number of edges.

By symmetry, it suffices to work under the assumption that $\omega'_e < \omega_e$. Clearly, this implies that $\alpha'_{h,n} \leq \alpha_{h,n}$ for all $h \in E_n$ and equality holds for $h = e$. Thus, by Lemma 4.3, we make the observation (II): every edge in T' other than e must also belong to T.

Let h be the unique edge that has weight $\alpha_{e,n}$. There are three possible scenarios: (a) If $\omega_h < \omega'_e < \omega_e$, then by Lemma 4.3, $e \notin T$ and $e \notin T'$. Therefore by the observations (I) and (II), $T = T'$. (b) If $\omega'_e < \omega_h < \omega_e$, then by Lemma 4.3, $e \in T'$, $h \notin T'$, $e \notin T$ and $h \in T$. By (I) and (II), this means that T' is obtained from T by deleting h and adding e. (c) If $\omega'_e < \omega_e < \omega_h$, then $e \in T$ and $e \in T'$, and therefore by (I) and (II), $T = T'$. In all three cases, it is easy to see that the formula for $\Delta_e M_n$ is valid. This completes the proof of Lemma 4.2. □

Lemma 4.2 gives an expression for $\Delta_e M_n$, but it does not make it obvious why this discrete difference is approximately equal to a local quantity. The secret lies in a monotonicity argument, similar in spirit to an idea from [53].

Lemma 4.4. *For any $e \in E$, the sequence $\alpha_{e,n}$ is a non-increasing sequence, converging everywhere to a limiting random variable $\alpha_{e,\infty}$ as $n \to \infty$. The convergence holds in L^p for every $p > 0$.*

Proof. The monotonicity is clear from the definition of $\alpha_{e,n}$. Since the sequence is non-negative, the limit exists. The L^p convergence holds because the random variables are bounded by a constant (since the edge weights are bounded by a constant). □

Now let c denote a specific edge of E, let's say the edge joining the origin to the point $(1, 0, \ldots, 0)$. For any edge e, let $e + V_n$ denote the set $x + [-n, n]^d \cap V_n$, where x is the lexicographically smaller endpoint of e. In other words, $e + V_n$ is simply a translate of V_n so that 0 maps to x. Let $e + E_n$ be the set of edges of $e + V_n$. For each e, let $\beta_{e,n}$ be the smallest β such that there is a path from one endpoint of e to the other, lying entirely in $e + V_n$ but not containing the edge e, such that all edges on this path have weight $\le \beta$. Clearly, $\beta_{e,n}$ has the same distribution as $\alpha_{c,n}$. The following lemma says that for a fixed edge e, if n and k and both large, and n is greater than k, then $\alpha_{e,n}$ may be closely approximated by $\beta_{e,k}$.

Lemma 4.5. *There is a sequence δ_k tending to zero as $k \to \infty$, such that for any $1 \le k < n$ and $e \in E_{n-k}$, $\mathbb{E}|\beta_{e,k} - \alpha_{e,n}| \le \delta_k$.*

Proof. Since $e + V_k \subseteq V_n$, $\beta_{e,k} \ge \alpha_{e,n}$. Thus, $\mathbb{E}|\beta_{e,k} - \alpha_{e,n}| = \mathbb{E}(\beta_{e,k}) - \mathbb{E}(\alpha_{e,n})$. But again, $V_n \subseteq e + V_{2n}$, and so $\alpha_{e,n} \ge \beta_{e,2n}$. Thus,

$$\mathbb{E}|\beta_{e,k} - \alpha_{e,n}| \le \mathbb{E}(\beta_{e,k}) - \mathbb{E}(\beta_{e,2n}) = \mathbb{E}(\alpha_{c,k}) - \mathbb{E}(\alpha_{c,2n})\,.$$

By Lemma 4.4, $\mathbb{E}(\alpha_{c,k})$ is a Cauchy sequence. This completes the proof. □

Combining Lemma 4.5 and Lemma 4.2, we get the following corollary that gives the desired "local approximation" for the discrete derivatives of M_n.

Corollary 4.6. *For any $k \ge 1$ and $e \in E$, let $\gamma_{e,k} := (\beta_{e,k} - \omega'_e)^+ - (\beta_{e,k} - \omega_e)^+$. Then for any $n > k$ and $e \in E_{n-k}$,*

$$\mathbb{E}|\Delta_e M_n - \gamma_{e,k}| \le 2\delta_k\,,$$

where δ_k is a sequence tending to zero as $k \to \infty$.

Armed with the above corollary and Corollary 3.2, we are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. Throughout this proof, C will denote any constant whose value depends only on the edge weight distribution and the dimension d. The value of C may change from line to line.

Fix an arbitrary positive integer k. Take any $n > k$. Take any edge $e \in E_{n-k}$, and a set of edges $A \subseteq E_n \backslash \{e\}$. Let $(\omega'_h)_{h \in E_n}$ be an independent copy of $(\omega_h)_{h \in E_n}$, and just like in Theorem 3.1, let $\omega_h^A = \omega_h$ if $h \notin A$, and $\omega_h^A = \omega'_h$ if $h \in A$. Let $\Delta_e M_n^A$ and $\gamma_{e,k}^A$ be the values of $\Delta_e M_n$ and $\gamma_{e,k}$ in the environment ω^A.

Let h be any other edge in E_{n-k} such that the lattice distance between e and h is bigger than $2k$. Let B be any subset of $E_n \backslash \{h\}$. Then by Corollary 4.6 and the boundedness of the discrete derivatives of M_n and the γ's, we get

$$|\mathrm{Cov}(\Delta_e M_n \Delta_e M_n^A,\, \Delta_h M_n \Delta_h M_n^B) - \mathrm{Cov}(\gamma_{e,k}\gamma_{e,k}^A,\, \gamma_{h,k}\gamma_{h,k}^B)| \le C\delta_k\,.$$

But since $(e + V_k) \cap (h + V_k) = \emptyset$, the random variables $\gamma_{e,k}\gamma_{e,k}^A$ and $\gamma_{h,k}\gamma_{h,k}^B$ are independent. In particular, their covariance is zero. Therefore,

$$|\mathrm{Cov}(\Delta_e M_n \Delta_e M_n^A,\, \Delta_h M_n \Delta_h M_n^B)| \le C\delta_k\,.$$

Note that here we are only considering e and h in E_{n-k} that are at least $2k$ apart in lattice distance. Therefore among all pairs of edges $e, h \in E_n$, we are excluding $\leq Cn^{2d-1}k$ pairs from the above bound. Those that are left out, are bounded by a constant.

All we now need is a lower bound on the variance σ_n^2. One can show that $\sigma_n^2 \geq Cn^d$. This requires some work, which is not necessary to present in this article. For a proof, see [29, Section 6.5]. Inputting this lower bound and the covariance bounds obtained in the above paragraph into Corollary 3.2, we get

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(f_n \leq t) - \mathbb{P}(Z \leq t)| \leq C(\delta_k + k/n)^{1/4} + Cn^{-d/4}\,.$$

The proof is finished by taking $n \to \infty$ and then taking $k \to \infty$. □

5. Some open problems

Probability theory has come a long way in figuring out how to prove central limit theorems. Still, there are problems where we do not know how to proceed. Many of these problems come from random combinatorial optimization. One example of a solved problem from this domain is the central limit theorem for minimal spanning trees, discussed in Section 4. But there are many others that are quite intractable.

For example, consider the Euclidean traveling salesman problem on a set of random points. Let $X_1, \ldots, X_n$ be a set of points chosen independently and uniformly at random from the unit square in $\mathbb{R}^2$. Let P be a path that visits all points, ending up where it started from, which minimizes the total distance traveled among all such paths. It is widely believed that the length of P should obey a central limit theorem under appropriate centering and scaling, but there is no proof.

Again, in the same setting, we may consider the problem of minimum matching. Suppose that n is even, and we pair the points into $n/2$ pairs such that the sum total of the pairwise distances is minimized. It is believed that this minimum matching length should be approximately normally distributed, but we do not know how to prove that.

One may also consider lattice versions of the above problems, where instead of points in Euclidean space we have random weights on the edges of a lattice. One can still talk about the minimum weight path that visits all points on a finite segment of the lattice, and the minimum weight matching of pairs of points. Central limit theorems should hold for both of these quantities.

For basic results about such models, a classic reference is the monograph of Steele [78]. The reason why one may speculate that normal approximation should hold is that the solutions of these problems are supposed to be "local" in nature. For example, the optimal path in the traveling salesman problem is thought to be of "locally determined"; one way to make this a little more precise is by claiming that a small perturbation at a particular location is unlikely to affect the path in some faraway neighborhood. This is the same as what we earlier called "the approximate independence of small perturbations". If this is proven to be indeed the case, then the generalized perturbative version of Stein's method should be an adequate tool for proving a central limit theorem.

Mean field versions of these problems, which look at complete graphs instead of lattices or Euclidean points, have been analyzed in great depth in a remarkable set of papers by Wästlund [85, 86]. In the case of minimum matching, this generalizes the famous work of

Aldous [1] on the random assignment problem. These papers, however, do not prove central limit theorems. It is an interesting question whether the insights gained from Wästlund's works can be applied to prove normal approximation in the mean field setting by rigorously proving the independence of small perturbations.

Another class of problems that may be attacked by high dimensional versions of Stein's method are problems of universality in physical models. There are various notions of universality; the one that is closest to standard probability theory is the following. Suppose that $Z = (Z_1, \ldots, Z_n)$ is a vector of i.i.d. standard normal random variables, and $X = (X_1, \ldots, X_n)$ is a vector of i.i.d. random variables from some other distribution, with mean zero and variance one. Let $f : \mathbb{R}^n \to \mathbb{R}$ be some given function. When is it true that $f(X)$ and $f(Z)$ have approximately the same probability distribution? In other words, when is it true that for all g belonging to a large class of functions, $\mathbb{E}g(f(X)) \approx \mathbb{E}g(f(Z))$? The classical central limit theorem says that this is true if $f(x) = n^{-1/2}(x_1 + \cdots + x_n)$. Lindeberg [54] gave an ingenious proof of the classical CLT in 1922 using the idea of replacing one X_i by one Z_i at a time, by an argument that I am going to describe below.

The idea was generalized by Rotar [77] to encompass low degree polynomials. The polynomial version was applied, in combination with hypercontractive estimates, to solve several open questions in theoretical computer science by Mossel et al. [58].

I think I may have been the first one to realize in [17, 19] that the Lindeberg method applies to general functions (and not just sums and polynomials), with a potentially wide range of interesting applications. The basic idea is the following: Let $h = g \circ f$. For each i, let $U^i = (X_1, \ldots, X_i, Z_{i+1}, \ldots, Z_n)$ and $V^i = (X_1, \ldots, X_{i-1}, 0, Z_{i+1}, \ldots, Z_n)$. Then by Taylor expansion in the ith coordinate,

$$\begin{aligned}\mathbb{E}h(U^i) - \mathbb{E}h(U^{i-1}) = \mathbb{E}\biggl(h(V^i) + X_i\partial_i h(V^i) + \frac{1}{2}X_i^2\partial_i^2 h(V^i)\biggr)\\ - \mathbb{E}\biggl(h(V^i) + Z_i\partial_i h(V^i) + \frac{1}{2}Z_i^2\partial_i^2 h(V^i)\biggr) + O(\|\partial_i^3 h\|_\infty)\,.\end{aligned}$$

By the independence of the X_i's and Z_i's, and the assumptions that $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$, it follows that the two expectations on the right-hand side are equal. Therefore, summing over i, we get

$$\mathbb{E}h(X) - \mathbb{E}h(Z) = O\biggl(\sum_{i=1}^n \|\partial_i^3 h\|_\infty\biggr)\,. \tag{5.1}$$

If the right-hand side is small, then we get our desired conclusion.

In [17, 19] I used this idea to give a new proof of the universality of Wigner's semicircle law, and a proof of the universality of the free energy of the Sherrington-Kirkpatrick model of spin glasses. The random matrix problems were tackled by choosing h to be the Stieltjes transform of the empirical spectral distribution of the random matrix at a point $z \in \mathbb{C}\backslash\mathbb{R}$. By taking z close to $\mathbb{R}$ and overcoming some major technical difficulties that arise in the process, the method was later used with great effect in a series of papers by Tao and Vu [82–84] to prove universality of local eigenvalue statistics of several kinds of random matrices.

The connection with Stein's method comes through the following variant of the Lindeberg idea. Suppose, instead of the above, we consider a solution w of the Stein equation

$$\Delta w(x) - x \cdot \nabla w(x) = h(x) - \mathbb{E}h(Z)\,.$$

Let $W^i := (X_1, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_n)$. Then by the independence of the X_i's and the facts that $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$, Taylor expansion gives

$$\begin{aligned}\mathbb{E}(X_i \partial_i w(X)) &= \mathbb{E}(X_i \partial_i w(W^i) + X_i^2 \partial_i^2 w(W^i)) + O(\|\partial_i^3 w\|_\infty) \\ &= \mathbb{E}\partial_i^2 w(W^i) + O(\|\partial_i^3 w\|_\infty) = \mathbb{E}\partial_i^2 w(X) + O(\|\partial_i^3 w\|_\infty)\,.\end{aligned}$$

Summing over i, this gives

$$\mathbb{E}h(X) - \mathbb{E}h(Z) = \mathbb{E}(\Delta w(X) - X \cdot \nabla w(X)) = O\biggl(\sum_{i=1}^n \|\partial_i^3 w\|_\infty\biggr)\,,$$

which is basically the same as (5.1), except that we have third derivatives of w instead of h. Undoubtedly, this is nothing but Stein's method in action. A version of this argument was used by Carmona and Hu [16] to prove the universality of the free energy in the Sherrington-Kirkpatrick model, at around the same time that I proved it in [17]. Sophisticated forms of this idea have been used by Erdős, Yau and coauthors in their remarkable series of papers [42–45] proving universality of random matrix eigenvalue distributions, running parallel to the papers of Tao and Vu, who used the Lindeberg approach. This demonstrates the potential for high dimensional versions of Stein's method to prove universality. There are still many problems where we do not know how to establish universal behavior (for example, last- and first-passage percolation, various polymer models, gradient Gibbs measures, etc.). It would be interesting to see Stein's method being used to attack such problems.

Acknowledgments. The author was partially supported by NSF grant DMS-1309618 during the preparation of this article. I thank Susan Holmes and Persi Diaconis for many useful comments on the first draft of this manuscript.

References

[1] Aldous, D. J., The $\zeta(2)$ limit in the random assignment problem. *Random Structures Algorithms,* **18** (2001), no. 4, 381–418.

[2] Alexander, K. S., *The RSW theorem for continuum percolation and the CLT for Euclidean minimal spanning trees*, Ann. Appl. Probab. **6** (1996), no. 2, 466–494.

[3] Arratia, R., Goldstein, L., and Gordon, L., *Two moments suffice for Poisson approximations: the Chen-Stein method*, Ann. Probab. **17** (1989), no. 1, 9–25.

[4] ———, *Poisson approximation and the Chen-Stein method*, Statist. Sci. **5** (1990), no. 4, 403–434.

[5] Avram, F. and Bertsimas, D., *On central limit theorems in geometrical probability*, Ann. Appl. Probab. **3** (1993), no. 4, 1033–1046.

[6] Baldi, P. and Rinott, Y., *On normal approximations of distributions in terms of dependency graphs*, Ann. Probab. **17** (1989), no. 4, 1646–1650.

[7] Baldi, P., Rinott, Y., and Stein, C., *A normal approximation for the number of local maxima of a random function on a graph*, In Probability, statistics, and mathematics pp. 59–81. Academic Press, Boston, MA, 1989.

[8] Barbour, A. D., *Stein's method for diffusion approximations*, Probab. Theory Related Fields **84** (1990), no. 3, 297–322.

[9] Barbour, A. D. and Chen, L. H. Y., editors, *An introduction to Stein's method*, Singapore University Press, Singapore, 2005.

[10] Barbour, A. D., Holst, L., and Janson, S., *Poisson approximation*, Oxford Science Publications, 1992.

[11] Barbour, A. D., Karoński, M., and Ruciński, A., *A central limit theorem for decomposable random variables with applications to random graphs*, J. Combin. Theory Ser. B **47** (1989), no. 2, 125–145.

[12] Bhattacharya, R. N. and Holmes, S. P., *An Exposition of Götze's Estimation of the Rate of Convergence in the Multivariate Central Limit Theorem*, In Bhattacharya, R. N. and Rao, R. R. (Eds.), Normal Approximation and Asymptotic Expansions (p. 260), SIAM, Philadelphia, PA, 2010.

[13] Bolthausen, E., *An estimate of the remainder in a combinatorial central limit theorem*, Probab. Theory Related Fields **66** (1984), no. 3, 379–386.

[14] Bolthausen, E. and Götze, F., *The rate of convergence for multivariate sampling statistics*, Ann. Statist. **21** (1993), 1692–1710.

[15] Borovkov, A. A. and Utev, S. A., *On an inequality and a related characterization of the normal distribution*, Theory Probab. Appl. **28**(2) (1984), 219–228.

[16] Carmona, P. and Hu, Y., *Universality in Sherrington-Kirkpatrick's spin glass model*, Ann. Inst. H. Poincaré Probab. Statist. **42** (2006), no. 2, 215–222.

[17] Chatterjee, S., *A simple invariance theorem* (2005), arXiv preprint.

[18] _____, *Concentration inequalities with exchangeable pairs*, Ph.D. dissertation, Stanford University, 2005.

[19] _____, *A generalization of the Lindeberg principle*, Ann. Probab. **34** (2006), no. 6, 2061–2076.

[20] _____, *Stein's method for concentration inequalities*, Probab. Theory Related Fields **138** (2007), nos. 1-2, 305–321.

[21] _____, *A new method of normal approximation*, Ann. Probab. **36** (2008), no. 4, 1584–1610.

[22] _____, *Fluctuations of eigenvalues and second order Poincaré inequalities*, Probab. Theory Related Fields **143** (2009) nos. 1-2, 1–40.

[23] _____, *Spin glasses and Stein's method*, Probab. Theory Related Fields **148** (2010), nos. 3-4, 567–600.

[24] _____, *A new approach to strong embeddings*, Probab. Theory Related Fields **152** (2012), nos. 1-2, 231–264.

[25] Chatterjee, S. and Dey, P. S., *Applications of Stein's method for concentration inequalities*, Ann. Probab. **38** (2010), no. 6, 2443–2485.

[26] Chatterjee, S., Diaconis, P., and Meckes, E., *Exchangeable pairs and Poisson approximation*, Probab. Surv. **2** (2005), 64–106.

[27] Chatterjee, S., Fulman, J., and Röllin, A., *Exponential approximation by Stein's method and spectral graph theory*, ALEA Lat. Am. J. Probab. Math. Stat. **8** (2011), 197–223.

[28] Chatterjee, S. and Meckes, E., *Multivariate normal approximation using exchangeable pairs*, ALEA Lat. Am. J. Probab. Math. Stat. **4** (2008), 257–283.

[29] Chatterjee, S. and Sen, S., *Minimal spanning trees and Stein's method* (2013), arXiv preprint.

[30] Chatterjee, S. and Shao, Q.-M., *Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model*, Ann. Appl. Probab. **21** (2011), no. 2, 464–483.

[31] Chatterjee, S. and Soundararajan, K., *Random multiplicative functions in short intervals*, Internat. Math. Research Notices **2012** (2012), no. 3, 479–492.

[32] Chen, L. H. Y., *Poisson approximation for dependent trials*, Ann. Probab. **3** (1975), no. 3, 534–545.

[33] ———, *The rate of convergence in a central limit theorem for dependent random variables with arbitrary index set*, IMA Preprint Series #243, Univ. Minnesota, 1986.

[34] ———, *The central limit theorem and Poincaré-type inequalities*, Ann. Probab. **16** (1988), no. 1, 300–304.

[35] Chen, L. H. Y., Fang, X., and Shao, Q.-M., *From Stein identities to moderate deviations*, Ann. Probab. **41** (2013), no. 1, 262–293.

[36] Chen, L. H. Y., Goldstein, L., and Shao, Q.-M., *Normal approximation by Stein's method*, Springer, Heidelberg, 2011.

[37] Chen, L. H. Y. and Röllin, A., *Stein couplings for normal approximation* (2010), arXiv preprint.

[38] Chen, L. H. Y. and Shao, Q.-M., *Normal approximation under local dependence*, Ann. Probab. **32** (2004), no. 3A, 1985–2028.

[39] Diaconis, P., *Stein's method for Markov chains: first examples*, In Stein's method: expository lectures and applications, 27–43, IMS Lecture Notes—Monograph Series, **46** (2004).

[40] Diaconis, P. and Holmes, S., editors, *Stein's method: expository lectures and applications*, IMS Lecture Notes—Monograph Series, **46** (2004).

[41] Eichelsbacher, P. and Löwe, M., *Stein's method for dependent random variables occurring in statistical mechanics*, Electron. J. Probab. **15** (2010), no. 30, 962–988.

[42] Erdős, L., Péché, S., Ramírez, J. A., Schlein, B., and Yau, H.-T., *Bulk universality for Wigner matrices*, Comm. Pure Appl. Math. **63** (2010), no. 7, 895–925.

[43] Erdős, L., Ramírez, J. A., Schlein, B., and Yau, H.-T., *Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation*, Electron. J. Probab. **15** (2010), no. 18, 526–603.

[44] Erdős, L., Schlein, B., and Yau, H.-T., *Universality of random matrices and local relaxation flow*, Invent. Math. **185** (2011), no. 1, 75–119.

[45] Erdős, L. and Yau, H.-T., *Universality of local spectral statistics of random matrices*, Bull. Amer. Math. Soc. (N.S.) **49** (2012), no. 3, 377–414.

[46] Fulman, J., *Stein's method and Plancherel measure of the symmetric group*, Trans. Amer. Math. Soc. **357** (2005), no. 2, 555–570.

[47] Ghosh, S. and Goldstein, L., *Applications of size biased couplings for concentration of measures*, Electr. Commun. Probab. **16** (2011a), 70–83.

[48] _____, *Concentration of measures via size-biased couplings*, Probab. Theory Related Fields **149** (2011b), 271–278.

[49] Goldstein, L. and Reinert, G., *Stein's method and the zero bias transformation with application to simple random sampling*, Ann. Appl. Probab. **7** (1997), no. 4, 935–952.

[50] Goldstein, L. and Rinott, Y., *Multivariate normal approximations by Stein's method and size bias couplings*, J. Appl. Probab. **33** (1996), no. 1, 1–17.

[51] Götze, F., *On the rate of convergence in the multivariate CLT*, Ann. Probab. **19** (1991), 724–739.

[52] Holmes, S., *Stein's method for birth and death chains*, In Stein's method: expository lectures and applications, 45–67, IMS Lecture Notes—Monogr. Ser. **46** (2004), Inst. Math. Statist., Beachwood, OH.

[53] Kesten, H. and Lee. S., *The central limit theorem for weighted minimal spanning trees on random points*, Ann. Appl. Probab. **6** (1996), no. 2, 495–527.

[54] Lindeberg, J. W., *Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung*, Math. Zeitschr. **15** (1922), 211–225.

[55] Luk, H. M., *Stein's method for the gamma distribution and related statistical applications*, Ph.D. thesis, University of Southern California, 1994.

[56] Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., and Tropp, J. A., *Matrix concentration inequalities via the method of exchangeable pairs* (2012), arXiv preprint.

[57] Mann, B., *Stein's method for χ^2 of a multinomial* (1997), Unpublished manuscript.

[58] Mossel, E., O'Donnell, R., and Oleszkiewicz, K., *Noise stability of functions with low influences: invariance and optimality*, Ann. of Math. (2) **171** (2010), no. 1, 295–341.

[59] Nolen, J., *Normal approximation for a random elliptic equation*, To appear in *Probab. Theory Related Fiels* (2011).

[60] Nourdin, I. and Peccati, G., *Stein's method on Wiener chaos*, Probab. Theory Related Fields **145** (2009), nos. 1-2, 75–118.

[61] ———, *Normal Approximations with Malliavin Calculus: From Stein's Method to Universality*, Cambridge University Press, 2012.

[62] Nourdin, I., Peccati, G., and Reinert, G., *Second order Poincaré inequalities and CLTs on Wiener space*, J. Funct. Anal. **257** (2009), no. 2, 593–609.

[63] Peköz, E. A., *Stein's method for geometric approximation*, J. Appl. Probab. **33** (1996), no. 3, 707–713.

[64] Peköz, E. A. and Röllin, A., *New rates for exponential approximation and the theorems of Rényi and Yaglom*, Ann. Probab. **39** (2011), no. 2, 587–608.

[65] Peköz, E. A., Röllin, A., and Ross, N., *Total variation error bounds for geometric approximation*, Bernoulli **19** (2013), no. 2, 610–632.

[66] Penrose, M. D., *Random geometric graphs*, Oxford University Press, Oxford, 2003.

[67] Raič, M., *CLT-related large deviation bounds based on Stein's method*, Adv. Appl. Probab. **39** (2007), no. 3, 731–752.

[68] Reinert, G., *Three general approaches to Stein's method*, In An introduction to Stein's method, volume 4 of Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap., pp. 183–221, Singapore Univ. Press, Singapore, 2005.

[69] Reinert, G. and Röllin, A., *Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition*, Ann. Probab. **37** (2009), no. 6, 2150–2173.

[70] ———, *Random subgraph counts and U-statistics: multivariate normal approximation via exchangeable pairs and embedding*, J. Appl. Probab. **47** (2010), no. 2, 378–393.

[71] Rinott, Y. and Rotar, V., *A multivariate CLT for local dependence with $n^{-1/2}\log n$ rate and applications to multivariate graph related statistics*, J. Multivariate Anal. **56** (1996), no. 2, 333–350.

[72] ———, *On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted U-statistics*, Ann. Appl. Probab. **7** (1997), no. 4, 1080–1105.

[73] ———, *On Edgeworth expansions for dependency-neighborhoods chain structures and Stein's method*, Probab. Theory Related Fields **126** (2003), no. 4, 528–570.

[74] Röllin, A., *A note on the exchangeability condition in Stein's method*, Statist. Probab. Lett. **78** (2008), no. 13, 1800–1806.

[75] ———, *Stein's method in high dimensions with applications*, Ann. Inst. Henri Poincaré (B): Probab. Stat. **49** (2013), no. 2, 529–549.

[76] Ross, N., *Fundamentals of Stein's method*, Probab. Surv. **8** (2011), 210–293.

[77] Rotar, V. I., *Limit theorems for polylinear forms*, J. Multivariate Anal. **9** (1979), 511–530.

[78] Steele, J. M., *Probability theory and combinatorial optimization*, SIAM, Philadelphia, PA, 1997.

[79] Stein, C., *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*, Proc. of the Sixth Berkeley Symp. on Math. Statist. and Probab., Vol. II: Probability theory (1972), 583–602.

[80] _____, *Approximate computation of expectations*, IMS Lecture Notes—Monograph Series, **7** (1986).

[81] Stein, C., Diaconis, P., Holmes, S., and Reinert, G., *Use of exchangeable pairs in the analysis of simulations*, In Stein's method: expository lectures and applications, IMS Lecture Notes–Monograph Series, **46** (2004), 1–26.

[82] Tao, T. and Vu, V., *Random matrices: universality of ESDs and the circular law. With an appendix by Manjunath Krishnapur*, Ann. Probab. **38** (2010a), no. 5, 2023–2065.

[83] _____, *Random matrices: universality of local eigenvalue statistics up to the edge*, Comm. Math. Phys. **298** (2010b), no. 2, 549–572.

[84] _____, *Random matrices: universality of local eigenvalue statistics*, Acta Math. **206** (2011), no. 1, 127–204.

[85] Wästlund, J., *The mean field traveling salesman and related problems*, Acta Math. **204** (2010), no. 1, 91–150.

[86] _____, *Replica symmetry of the minimum matching*, Ann. of Math. (2) **175** (2012), no. 3, 1061–1091.

Department of Statistics and Department of Mathematics, Stanford University, USA
E-mail: souravc@stanford.edu

Criticality, universality, and isoradiality

Geoffrey R. Grimmett

Abstract. Critical points and singularities are encountered in the study of critical phenomena in probability and physics. We present recent results concerning the values of such critical points and the nature of the singularities for two prominent probabilistic models, namely percolation and the more general random-cluster model. The main topic is the statement and proof of the criticality and universality of the canonical measure of bond percolation on isoradial graphs (due to the author and Ioan Manolescu). The key technique used in this work is the star–triangle transformation, known also as the Yang–Baxter equation. The second topic reported here is the identification of the critical point of the random-cluster model on the square lattice (due to Beffara and Duminil-Copin), and of the criticality of the canonical measure of the random-cluster model with $q \geq 4$ on periodic isoradial graphs (by the same authors with Smirnov). The proof of universality for percolation is expected to extend to the random-cluster model on isoradial graphs.

Mathematics Subject Classification (2010). Primary 60K35; Secondary 82B20.

Keywords. Percolation, random-cluster model, Ising/Potts models, critical point, universality, isoradial graph, critical exponent, star–triangle transformation, Yang–Baxter equation.

1. Introduction

One of the most provocative and elusive problems in the mathematics of critical phenomena is the issue of *universality*. Disordered physical systems manifest phase transitions, the nature of which is believed to be independent of the local structure of space. Very little about universality is known rigorously for systems below their upper critical dimension. It is frequently said that "renormalization" is the key to universality, but rigorous applications of renormalization in the context of universality are rare.

There has been serious recent progress in the "exactly solvable" setting of the two-dimensional Ising model, and a handful of special cases for other models. Our principal purpose here is to outline recent progress concerning the identification of critical surfaces and the issue of universality for bond percolation and the random-cluster model on isoradial graphs, with emphasis on the general method, the current limitations, and the open problems.

For bond percolation on an extensive family of isoradial graphs, the canonical process, in which the star–triangle transformation is in harmony with the geometry, is shown to be critical. Furthermore, universality has been proved for this class of systems, at least for the critical exponents *at the critical surface*. These results, found in recent papers by the author and Manolescu, [27–29], vastly extend earlier calculations of critical values for the square lattice etc, with the added ingredient of universality. Note that, to date, we are able to prove

Proceedings of the International Congress of Mathematicians, Seoul, 2014

only *conditional* universality: if a certain exponent exists for at least one isoradial graph, then a family of exponents exist for an extensive collection of isoradial graphs, and they are universal across this collection.

The picture for the general random-cluster model is more restrained, but significant progress has been achieved on the identification of critical points. The longstanding conjecture for the critical value of the square lattice has been proved by Beffara and Duminil-Copin [4], using a development of classical tools. Jointly with Smirnov [5], the same authors have used Smirnov's parafermionic observable in the first-order setting of $q \geq 4$ to identify the critical surface of a periodic isoradial graph. It is conjectured that the methods of [29] may be extended to obtain universality for the random-cluster model on isoradial graphs.

The results reported in this survey are closely related to certain famous 'exact results' in the physics literature. Prominent in the latter regard is the book of Baxter [3], from whose preface we quote selectively as follows:

> "... the phrase 'exactly solved' has been chosen with care. It is not necessarily the same as 'rigorously solved'. ... There is of course still much to be done."

Percolation is summarized in Section 2, and isoradial graphs in Section 3. Progress with criticality and universality for percolation are described in Section 4. Section 6 is devoted to recent progress with critical surfaces of random-cluster models on isoradial graphs, and open problems for percolation and the random-cluster model may be found in Sections 5 and 7.

2. Percolation

2.1. Background. Percolation is the fundamental stochastic model for spatial disorder. Since its introduction by Broadbent and Hammersley in 1957, it has emerged as a key topic in probability theory, with connections and impact across all areas of applied science in which disorder meets geometry. It is in addition a source of beautiful and apparently difficult mathematical problems, the solutions to which often require the development of new tools with broader applications.

Here is the percolation process in its basic form. Let $G = (V, E)$ be an infinite, connected graph, typically a crystalline lattice such as the d-dimensional hypercubic lattice. We are provided with a coin that shows heads with some fixed probability p. For each edge e of G, we flip the coin, and we designate e *open* if heads shows, and *closed* otherwise. The open edges are considered open to the passage of material such as liquid, disease, or rumour.[1]

Liquid is supplied at a *source* vertex s, and it flows along the open edges and is blocked by the closed edges. The basic problem is to determine the geometrical properties (such as size, shape, and so on) of the region C_s that is wetted by the liquid. More generally, one is interested in the geometry of the connected subgraphs of G induced by the set of open edges. The components of this graph are called the *open clusters*.

Broadbent and Hammersley proved in [10, 30, 31] that there exists a *critical probability* $p_c = p_c(G)$ such that: every open cluster is bounded if $p < p_c$, and some open cluster is unbounded if $p > p_c$. There are two *phases*: the *subcritical phase* when $p < p_c$ and the

[1] This is the process known as *bond* percolation. Later we shall refer to *site* percolation, in which the vertices (rather than the edges) receive random states.

supercritical phase when $p > p_c$. The singularity that occurs when p is near or equal to p_c has attracted a great deal of attention from mathematicians and physicists, and many of the principal problems remain unsolved even after several decades of study. See [22, 25] for general accounts of the theory of percolation.

Percolation is one of a large family of models of classical and quantum statistical physics that manifest phase transitions, and its theory is near the heart of the extensive scientific project to understand phase transitions and critical phenomena. Key aspects of its special position in the general theory include: (i) its deceptively simple formulation as a probabilistic model, (ii) its use as a comparator for more complicated systems, and (iii) its role in the development of new methodology.

One concrete connection between percolation and models for ferromagnetism is its membership of the one-parameter family of so-called random-cluster models. That is, percolation is the $q = 1$ random-cluster model. The $q = 2$ random-cluster model corresponds to the Ising model, and the $q = 3, 4, \dots$ random-cluster models to the q-state Potts models. The $q \downarrow 0$ limit is connected to electrical networks, uniform spanning trees, and uniform connected subgraphs. The *geometry* of the random-cluster model corresponds to the *correlation* structure of the Ising/Potts models, and thus its critical point p_c may be expressed in terms of the critical temperature of the latter systems. See [23, 64] for a general account of the random-cluster model.

The theory of percolation is extensive and influential. Not only is percolation a benchmark model for studying random spatial processes in general, but also it has been, and continues to be, a source of intriguing and beautiful open problems. Percolation in two dimensions has been especially prominent in the last decade by virtue of its connections to conformal invariance and conformal field theory. Interested readers are referred to the papers [14, 26, 54, 56, 57, 61, 63] and the books [6, 22, 25].

2.2. Formalities. For $x, y \in V$, we write $x \leftrightarrow y$ if there exists an open path joining x and y. The *open cluster* at the vertex x is the set $C_x = \{y : x \leftrightarrow y\}$ of all vertices reached along open paths from x, and we write $C = C_0$ where 0 is a fixed vertex called the *origin*. Write $\mathbb{P}_p$ for the relevant product probability measure, and $\mathbb{E}_p$ for expectation with respect to $\mathbb{P}_p$.

The *percolation probability* is the function $\theta(p)$ given by

$$\theta(p) = \mathbb{P}_p(|C| = \infty),$$

and the *critical probability* is defined by

$$p_c = p_c(G) = \sup\{p : \theta(p) = 0\}. \tag{2.1}$$

It is elementary that θ is a non-decreasing function, and therefore,

$$\theta(p) \begin{cases} = 0 & \text{if } p < p_c, \\ > 0 & \text{if } p > p_c. \end{cases}$$

It is not hard to see, by the Harris–FKG inequality, that the value $p_c(G)$ does not depend on the choice of origin.

Let $d \geq 2$, and let $\mathcal{L}$ be a d-dimensional lattice. It is a fundamental fact that $0 < p_c(\mathcal{L}) < 1$, but it is unproven in general that no infinite open cluster exists when $p = p_c$

Conjecture 2.1. *For any lattice $\mathcal{L}$ in $d \geq 2$ dimensions, we have that $\theta(p_c) = 0$.*

The claim of the conjecture is known to be valid for certain lattices when $d = 2$ and for large d, currently $d \geq 15$. This conjecture has been the 'next open problem' since the intensive study of the late 1980s.

Whereas the above process is defined in terms of a single parameter p, we are concerned here with the richer multi-parameter setting in which an edge e is designated open with some probability p_e. In such a case, the critical probability p_c is replaced by a so-called 'critical surface'.

2.3. Critical exponents and universality. A great deal of effort has been directed towards understanding the nature of the percolation phase transition. The picture is now fairly clear for one specific model in two dimensions (site percolation on the triangular lattice), owing to the very significant progress in recent years linking critical percolation to the Schramm–Löwner curve SLE_6. There remain however substantial difficulties to be overcome even when $d = 2$, associated largely with the extension of such results to general two-dimensional systems. The case of large d (currently, $d \geq 15$) is also well understood, through work based on the so-called 'lace expansion' (see [1]). Many problems remain open in the prominent case $d = 3$.

Let $\mathcal{L}$ be a d-dimensional lattice. The nature of the percolation singularity on $\mathcal{L}$ is expected to share general features with phase transitions of other models of statistical mechanics. These features are sometimes referred to as 'scaling theory' and they relate to the 'critical exponents' occurring in the power-law singularities (see [22, Chap. 9]). There are two sets of critical exponents, arising firstly in the limit as $p \to p_c$, and secondly in the limit over increasing spatial scales when $p = p_c$. The definitions of the critical exponents are found in Table 2.1 (taken from [22]).

The notation of Table 2.1 is as follows. We write $f(x) \approx g(x)$ as $x \to x_0 \in [0, \infty]$ if $\log f(x)/\log g(x) \to 1$. The *radius* of the open cluster C at the origin x is defined by

$$\mathrm{rad}(C) = \sup\{\|y\| : x \leftrightarrow y\},$$

where

$$\|y\| = \sup_i |y_i|, \qquad y = (y_1, y_2, \ldots, y_d) \in \mathbb{R}^d,$$

is the supremum (L^∞) norm on $\mathbb{R}^d$. The limit as $p \to p_c$ should be interpreted in a manner appropriate for the function in question (for example, as $p \downarrow p_c$ for $\theta(p)$, but as $p \to p_c$ for $\kappa(p)$). The *indicator function* of an event A is denoted 1_A.

Eight critical exponents are listed in Table 2.1, denoted α, β, γ, δ, ν, η, ρ, Δ, but there is no general proof of the existence of any of these exponents for arbitrary $d \geq 2$. Such critical exponents may be defined for phase transitions in a large family of physical systems. The exponents are not believed to be independent variables, but rather to satisfy the so-called *scaling relations*

$$2 - \alpha = \gamma + 2\beta = \beta(\delta + 1),$$
$$\Delta = \delta\beta, \quad \gamma = \nu(2 - \eta),$$

and, when d is not too large, the *hyperscaling relations*

$$d\rho = \delta + 1, \quad 2 - \alpha = d\nu.$$

More generally, a 'scaling relation' is any equation involving critical exponents which is believed to be 'universally' valid. The *upper critical dimension* is the largest value d_c such

Function		*Behaviour*	*Exp.*
percolation probability	$\theta(p) = \mathbb{P}_p(\lvert C\rvert = \infty)$	$\theta(p) \approx (p - p_c)^\beta$	β
truncated mean cluster-size	$\chi^f(p) = \mathbb{E}_p(\lvert C\rvert 1_{\lvert C\rvert<\infty})$	$\chi^f(p) \approx \lvert p - p_c\rvert^{-\gamma}$	γ
number of clusters per vertex	$\kappa(p) = \mathbb{E}_p(\lvert C\rvert^{-1})$	$\kappa'''(p) \approx \lvert p - p_c\rvert^{-1-\alpha}$	α
cluster moments	$\chi^f_k(p) = \mathbb{E}_p(\lvert C\rvert^k 1_{\lvert C\rvert<\infty})$	$\dfrac{\chi^f_{k+1}(p)}{\chi^f_k(p)} \approx \lvert p - p_c\rvert^{-\Delta}$	Δ
correlation length	$\xi(p)$	$\xi(p) \approx \lvert p - p_c\rvert^{-\nu}$	ν
cluster volume		$\mathbb{P}_{p_c}(\lvert C\rvert = n) \approx n^{-1-1/\delta}$	δ
cluster radius		$\mathbb{P}_{p_c}\big(\mathrm{rad}(C) = n\big) \approx n^{-1-1/\rho}$	ρ
connectivity function		$\mathbb{P}_{p_c}(0 \leftrightarrow x) \approx \lVert x\rVert^{2-d-\eta}$	η

Table 2.1. Eight functions and their critical exponents. The first five exponents arise in the limit as $p \to p_c$, and the remaining three as $n \to \infty$ with $p = p_c$. See [22, p. 127] for a definition of the correlation length $\xi(p)$.

that the hyperscaling relations hold for $d \le d_c$ and not otherwise. It is believed that $d_c = 6$ for percolation. There is no general proof of the validity of the scaling and hyperscaling relations for percolation, although quite a lot is known when either $d = 2$ or d is large. The case of large d is studied via the lace expansion, and this is expected to be valid for $d > 6$.

We note some further points in the context of percolation.

(a) **Universality**. The numerical values of critical exponents are believed to depend only on the value of d, and to be independent of the choice of lattice, and of the type of percolation under study.

(b) **Two dimensions**. When $d = 2$, it is believed that

$$\alpha = -\tfrac{2}{3},\ \beta = \tfrac{5}{36},\ \gamma = \tfrac{43}{18},\ \delta = \tfrac{91}{5}, \dots.$$

These values (other than α) have been proved (essentially only) in the special case of site percolation on the triangular lattice, see [45, 60].

(c) **Large dimensions**. When d is sufficiently large (in fact, $d \ge d_c$) it is believed that the critical exponents are the same as those for percolation on a tree (the 'mean-field model'), namely $\delta = 2$, $\gamma = 1$, $\nu = \frac{1}{2}$, $\rho = \frac{1}{2}$, and so on. Using the first hyperscaling

relation, this is consistent with the contention that $d_c = 6$. Several such statements are known to hold for $d \geq 15$, see [20, 32, 33, 41].

Open challenges include the following:

1. prove the existence of critical exponents for general lattices,
2. prove some version of universality,
3. prove the scaling and hyperscaling relations in general dimensions,
4. calculate the critical exponents for general models in two dimensions,
5. prove the mean-field values of critical exponents when $d \geq 6$.

Progress towards these goals has been substantial but patchy. As noted above, for sufficiently large d, the lace expansion has enabled proofs of exact values for many exponents, for a restricted class of lattices. There has been remarkable progress in recent years when $d = 2$, inspired largely by work of Cardy [14] and Schramm [53], enacted by Smirnov [56], and confirmed by the programme pursued by Lawler, Schramm, Werner, Camia, Newman, Sheffield and others to understand SLE curves and conformal ensembles.

In this paper, we concentrate on recent progress concerning isoradial embeddings of planar graphs, and particularly the identification of their critical surfaces and the issue of universality.

3. Isoradial graphs

Let G be an infinite, planar graph embedded in $\mathbb{R}^2$ in such a way that edges intersect only at vertices. For simplicity, we assume that the embedding has only bounded faces. The graph G is called *isoradial* if (i) every face has a circumcircle which passes through every vertex of the face, (ii) the centre of each circumcircle lies in the interior of the corresponding face, and (iii) all such circumcircles have the same radius. We may assume by re-scaling that the common radius is 1.

The family of isoradial graphs is in two-to-one correspondence with the family of tilings of the plane with rhombi of side-length 1, in the following sense. Consider a rhombic tiling of the plane, as in Figure 3.1. The tiling, when viewed as a graph, is bipartite with vertex-sets coloured red and white, say. Fix a colour and join any two vertices of that colour whenever they are the opposite vertices of a rhombus. The resulting graph G is isoradial. If the other colour is chosen, the resulting graph is the (isoradial) dual of G. This is illustrated in Figures 3.1 and 3.2. Conversely, given an isoradial graph G, the corresponding rhombic tiling is obtained by augmenting its vertex-set by the circumcentres of the faces, and each circumcentre is joined to the vertices of the enclosing face.

Isoradial graphs were introduced by Duffin [17], and are related to the so-called Z-invariant graphs of Baxter [2]. They were named thus by Kenyon, whose expository paper [36] proposes the connection between percolation and isoradiality (and much more). Isoradial graphs have two important properties, the first of which is their connection to preholomorphic functions. This was discovered by Duffin, and is summarized by Smirnov [59] and developed further in the context of probability by Chelkak and Smirnov [15]. This property is key to the work on the random-cluster model on isoradial graphs reviewed in Section

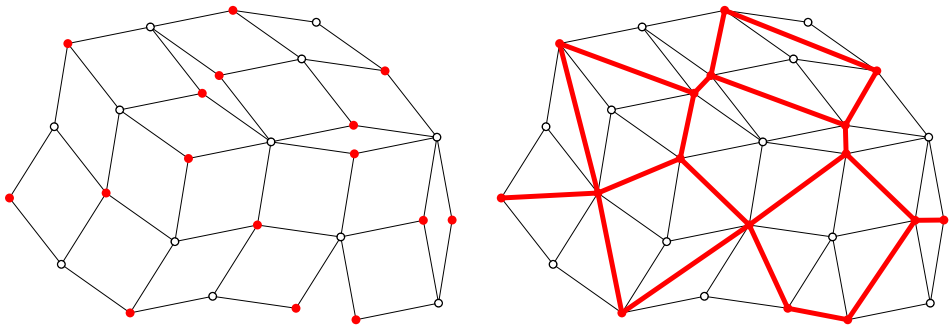


Figure 3.1. On the left is part of a rhombic tiling of the plane. Since all cycles have even length, this is a bipartite graph, with vertex-sets coloured red and white. The graph on the right is obtained by joining pairs of red vertices across faces. Each red face of the latter graph contains a unique white vertex, and this is the centre of the circumcircle of that face. Joining the white vertices, instead, yields another isoradial graph that is dual to the first.

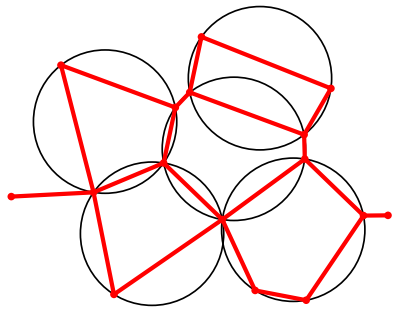


Figure 3.2. An illustration of the isoradiality of the red graph of Figure 3.1.

6. A recent review of connections between isoradiality and aspects of statistical mechanics may be found in [8].

The second property of isoradial graphs is of special relevance in the current work, namely that they provide the 'right' setting for the star–triangle transformation. This is explained next.

Consider an inhomogeneous bond percolation process on the isoradial graph G, whose edge-probabilities p_e are given as follows in terms of the graph-embedding. Each edge e of G is the diagonal of a unique rhombus in the corresponding rhombic tiling of the plane, and its parameter p_e is given in terms of the geometry of this rhombus. With θ_e the opposite angle of the rhombus, as illustrated in Figure 3.3, let $p_e \in (0,1)$ satisfy

$$\frac{p_e}{1-p_e} = \frac{\sin(\frac{1}{3}[\pi - \theta_e])}{\sin(\frac{1}{3}\theta_e)}. \tag{3.1}$$

We consider inhomogeneous bond percolation on G in which each edge e is designated open with probability p_e, and we refer to this as the *canonical percolation process* on G, with associated probability measure $\mathbb{P}_G$. The special property of the vector $\mathbf{p} = (p_e : e \in E)$ is explained in Section 4.2.

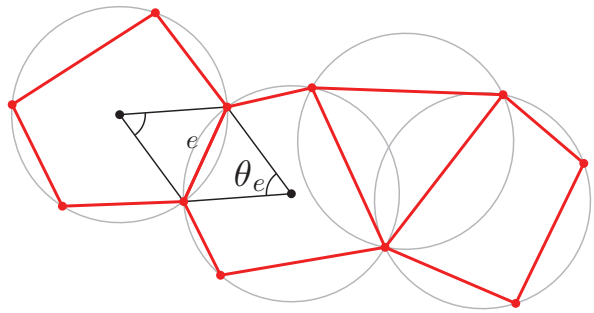


Figure 3.3. The edge e is the diagonal of some rhombus, with opposite angle θ_e as illustrated.

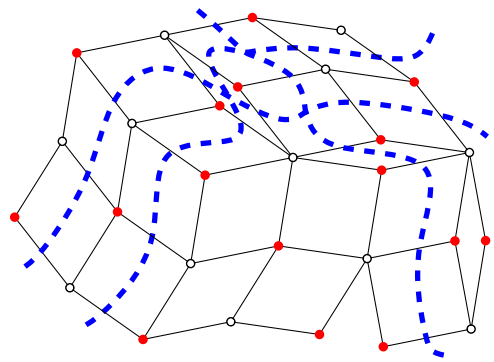


Figure 3.4. An illustration of the track system of the rhombic tiling of Figure 3.1.

In a beautiful series of papers [11–13], de Bruijn introduced the geometrical construct of 'ribbons' or 'train tracks' via which he was able to build a theory of rhombic tilings. Consider a tiling $\mathbb{T}$ of the plane in which each tile is convex with four sides. We pursue a walk on the faces of $\mathbb{T}$ according to the following rules. The walk starts in some given tile, and crosses some edge to a neighbouring tile. It next traverses the opposite edge of this tile, and so on. The walk may be extended backwards according to the same rule, and a doubly-infinite walk ensues. Such a walk is called a *ribbon* or *track*. De Bruijn pointed out that, if $\mathbb{T}$ is a rhombic tiling, then no walk intersects itself, and two walks may intersect once but not twice. This property turns out to be both necessary and sufficient for a track system to be homeomorphic to that of a *rhombic* tiling (see [37]).

We impose two restrictions on the isoradial graphs under study. Firstly, we say that an isoradial graph $G = (V, E)$ satisfies the *bounded-angles property* (BAP) if there exists $\epsilon > 0$ such that

$$\epsilon < \theta_e < \pi - \epsilon \qquad \text{for all } e \in E,$$

where θ_e is as in Figure 3.3. This amounts to the condition that the rhombi in the corresponding tiling are not 'too flat'. We say that G has the *square-grid property* (SGP) if its track system, viewed as a graph, contains a square grid such that those tracks not in the grid have boundedly many intersections with the grid within any bounded region (see [29, Sect. 4.2] for a more careful statement of this property).

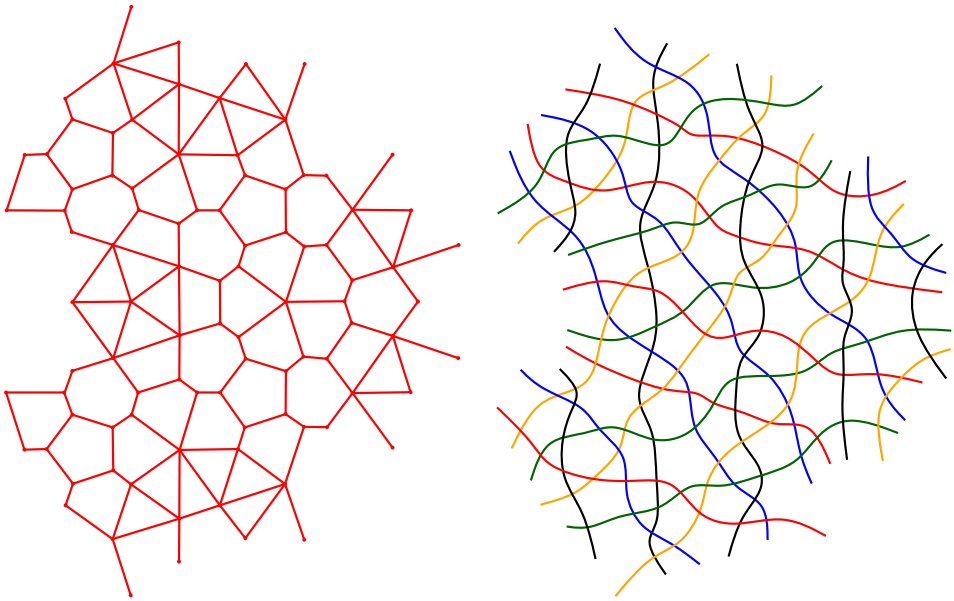


Figure 3.5. On the left, an isoradial graph obtained from part of the Penrose rhombic tiling. On the right, the associated track system comprises a pentagrid: five sets of non-intersecting doubly-infinite lines.

An isoradial graph may be viewed as both a graph and a planar embedding of a graph. Of the many examples of isoradial graphs, we mention first the conventional embeddings of the square, triangular, and hexagonal lattices. These are symmetric embeddings, and the edges have the same p-value. There are also non-symmetric isoradial embeddings of the same lattices, and indeed embeddings with no non-trivial symmetry, for which the corresponding percolation measures are 'highly inhomogeneous'.

The isoradial family is much richer than the above examples might indicate, and includes graphs obtained from aperiodic tilings including the classic Penrose tiling [49, 50], illustrated in Figure 3.5. All isoradial graphs mentioned above satisfy the SGP, and also the BAP so long as the associated tiling comprises rhombi with flatness uniformly bounded from 0.

4. Criticality and universality for percolation

4.1. Two main theorems. The first main theorem of [29] is the identification of the criticality of the canonical percolation measure $\mathbb{P}_G$ on an isoradial graph G. The second is the universality of $\mathbb{P}_G$ across an extensive family of isoradial graphs G.

In order to state the criticality theorem, we introduce notation that is appropriate for a perturbation of the canonical measure $\mathbb{P}_G$, and we borrow that of [5]. For $e \in E$ and $\beta \in (0, \infty)$, let $p_e(\beta)$ satisfy

$$\frac{p_e(\beta)}{1 - p_e(\beta)} = \beta \frac{\sin(\frac{1}{3}[\pi - \theta_e])}{\sin(\frac{1}{3}\theta_e)}, \tag{4.1}$$

and write $\mathbb{P}_{G,\beta}$ for the corresponding product measure on G. Thus $\mathbb{P}_{G,1} = \mathbb{P}_G$.

Theorem 4.1 (Criticality [29]). *Let $G = (V, E)$ be an isoradial graph with the bounded-angles property and the square-grid property. The canonical percolation measure $\mathbb{P}_G$ is critical in that*

(a) *there exist $a, b, c, d > 0$ such that*

$$ak^{-b} \le \mathbb{P}_G\big(\mathrm{rad}(C_v) \ge k\big) \le ck^{-d}, \qquad k \ge 1, \quad v \in V,$$

(b) *there exists, $\mathbb{P}_G$-a.s., no infinite open cluster,*

(c) *for $\beta < 1$, there exist $f, g > 0$ such that*

$$\mathbb{P}_{G,\beta}(|C_v| \ge k) \le fe^{-gk}, \qquad k \ge 0, \quad v \in V,$$

(d) *for $\beta > 1$, there exists, $\mathbb{P}_{G,\beta}$-a.s., a unique infinite open cluster.*

This theorem includes as special cases a number of known results for homogeneous and inhomogenous percolation on the square, triangular, and hexagonal lattices beginning with Kesten's theorem that $p_c = \frac{1}{2}$ for the square lattice, see [38, 39, 65].

We turn now to the universality of critical exponents. Recall the exponents ρ, η, and δ of Table 2.1. The exponent ρ_{2j} is the so-called $2j$ alternating-arm critical exponent, see [26, 29]. An exponent is said to be *$\mathcal{G}$-invariant* if its value is constant across the family $\mathcal{G}$.

Theorem 4.2 (Universality [29]). *Let $\mathcal{G}$ be the class of isoradial graphs with the bounded-angles property and the square-grid property.*

(a) *Let $\pi \in \{\rho\} \cup \{\rho_{2j} : j \ge 1\}$. If π exists for some $G \in \mathcal{G}$, then it is $\mathcal{G}$-invariant.*

(b) *If either ρ or η exists for some $G \in \mathcal{G}$, then ρ, η, δ are $\mathcal{G}$-invariant and satisfy the scaling relations $\eta\rho = 2$ and $2\rho = \delta + 1$.*

The theorem establishes universality for bond percolation on isoradial graphs, but restricted to the exponents ρ, η, δ *at the critical point*. The method of proof does not seem to extend to the near-critical exponents β, γ, etc (see Problem E of Section 5).

It is in fact 'known' that, for reasonable two-dimensional lattices,

$$\rho = \tfrac{48}{5}, \quad \eta = \tfrac{5}{24}, \quad \delta = \tfrac{91}{5}, \tag{4.2}$$

although these values (and more), long predicted in the physics literature, have been proved rigorously only for (essentially) site percolation on the triangular lattice. See Lawler, Schramm, Werner [45] and Smirnov and Werner [60]. Note that site percolation on the triangular lattice does not lie within the ambit of Theorems 4.1 and 4.2.

To summarize, there is currently no known proof of the existence of critical exponents for any graph belonging to $\mathcal{G}$. However, if certain exponents exist for *any* such graph, then they exist for all G and are $\mathcal{G}$-invariant. If one could establish a result such as in (4.2) for any such graph, then this result would be valid across the entire family $\mathcal{G}$.

The main ideas of the proofs of Theorems 4.1 and 4.2 are as follows. The first element is the so-called box-crossing property. Loosely speaking, this is the property that the probability of an open crossing of a box with given aspect-ratio is bounded away from 0, uniformly in the position, orientation, and size of the box. The box-crossing property was proved by

Russo [52] and Seymour/Welsh [55] for homogeneous percolation on the square lattice, using its properties of symmetry and self-duality. It may be shown using classical methods that the box-crossing property is a certificate of a critical or supercritical percolation model. It may be deduced that, if both the primal and dual models have the box-crossing property, then they are both critical.

The star–triangle transformation of the next section provides a method for transforming one isoradial graph into another. The key step in the proofs is to show that this transformation preserves the box-crossing property. It follows that any isoradial graph that can be obtained by a sequence of transformations from the square lattice has the box-crossing property, and is therefore critical. It is proved in [29] that this includes any isoradial graph with both the BAP and SGP.

4.2. Star–triangle transformation. The central fact that permits proofs of criticality and universality is that the star–triangle transformation has a geometric representation that acts locally on rhombic tilings. Consider three rhombi assembled to form a hexagon as in the upper left of Figure 4.1. The interior of the hexagon may be tiled by (three) rhombi in either of two ways, the other such tiling being drawn at the upper right of the figure. The switch from the first to the second has two effects: (i) the track system is altered as indicated there, with one track being moved over the intersection of the other two, and (ii) the triangle in the isoradial graph of the upper left is transformed into a star. These observations are graph-theoretic rather than model-specific. We next parametrize the system in such a way that the parameters mutate in the canonical way under the above transformation. That is, for a given probabilistic model, we seek a parametrization under which the geometrical switch induces the appropriate parametric change.

Here is the star–triangle transformation for percolation. Consider the triangle $T = (V, E)$ and the star $S = (V', E')$ as drawn in Figure 4.2. Let $\mathbf{p} = (p_0, p_1, p_2) \in [0, 1)^3$, and suppose the edges in the figure are declared open with the stated probabilities. The two ensuing configurations induce two connectivity relations on the set $\{A, B, C\}$ within S and T, respectively. It turns out that these two connectivity relations are equi-distributed so long as $\kappa(\mathbf{p}) = 0$, where

$$\kappa(\mathbf{p}) = p_0 + p_1 + p_2 - p_1p_2p_3 - 1. \tag{4.3}$$

The star–triangle transformation is used as follows. Suppose, in a graph G, one finds a triangle whose edge-probabilities satisfy (4.3). Then this triangle may be replaced by a star having the complementary probabilities of Figure 4.2 without altering the probabilities of any long-range connections in G. Similarly, stars may be transformed into triangles. One complicating feature of the transformation is the creation of a new vertex when passing from a triangle to a star (and its destruction when passing in the reverse direction).

The star–triangle transformation was discovered first in the context of electrical networks by Kennelly [35] in 1899, and it was adapted in 1944 by Onsager [48] to the Ising model in conjunction with Kramers–Wannier duality. It is a key element in the work of Baxter [2, 3] on exactly solvable models in statistical mechanics, where it has become known as the *Yang–Baxter equation* (see [51] for a history of its importance in physics). Sykes and Essam [62] used the star–triangle transformation to predict the critical surfaces of certain inhomogeneous (but periodic) bond percolation processes on the triangular and hexagonal lattices, and furthermore the star–triangle transformation is a tool in the study of the random-cluster model [23, Sect. 6.6], and the dimer model [7].

Let us now explore the operation of the star–triangle transformation in the context of the

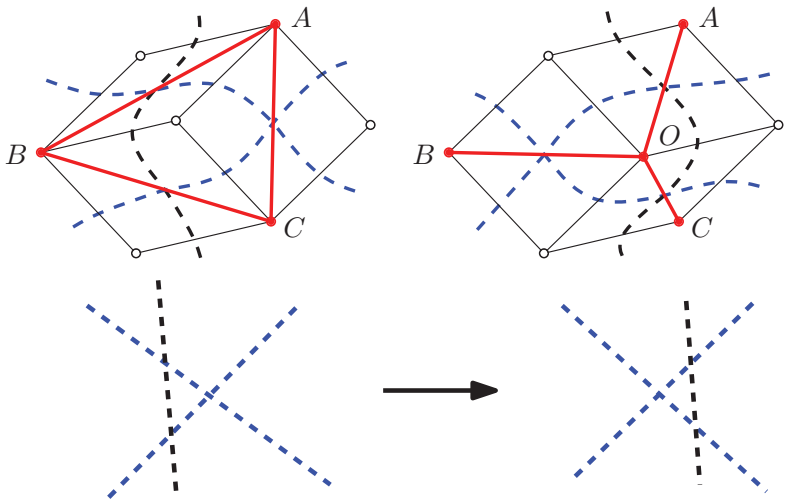


Figure 4.1. There are two ways of tiling the hexagon in the upper figure, and switching between these amounts to a star–triangle transformation for the isoradial graph. The effect on the track system is illustrated in the lower figure.

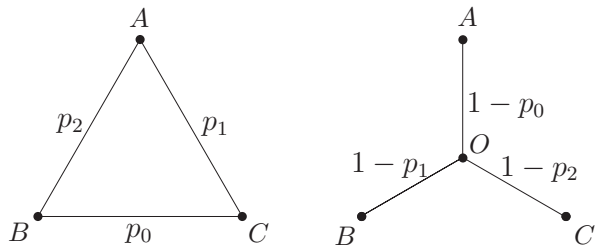


Figure 4.2. The star–triangle transformation for bond percolation.

rhombic switch of Figure 4.1. Let G be an isoradial graph containing the upper left hexagon of the figure, and let G' be the new graph after the rhombic switch. The definition (3.1) of the edge-probabilities has been chosen in such a way that the values on the triangle satisfy (4.3) and those on the star are as given in Figure 4.2. It follows that the connection probabilities on G and G' are equal. Graphs which have been thus parametrized but not embedded isoradially were called *Z-invariant* by Baxter [2]. See [44] for a recent account of the application of the above rhombic switch to Glauber dynamics of lozenge tilings of the triangular lattice.

One may couple the probability spaces on G and G' in such a way that the star–triangle transformation preserves *open connections*, rather than just their probabilities. Suppose that, in a given configuration, there exists an open path in G between vertex-sets A and B. On applying a sequence of star–triangle transformations, we obtain an open path in G' from the image of A to the image of B. Thus, star–triangle transformations transport open paths to open paths, and it is thus that the box-crossing property is transported from G to G'.

In practice, infinitely many star–triangle transformations are required to achieve the necessary transitions between graphs. The difficulties of the proofs of Theorems 4.1–4.2 are

centred on the need to establish sufficient control on the drifts of paths and their endvertices under these transformations.

5. Open problems for percolation

We discuss associated open problems in this section.

(A) **Existence and equality of critical exponents.** It is proved in Theorem 4.2 that, if the three exponents ρ, η, δ exist for some member of the family $\mathcal{G}$, then they exist for all members of the family, and are constant across the family. Essentially the only model for which existence has been proved is the site model on the triangular lattice, but this does not belong to $\mathcal{G}$. A proof of existence of exponents for the bond model on the square lattice would imply their existence for the isoradial graphs studied here. Similarly, if one can show any exact value for the latter bond model, then this value holds across $\mathcal{G}$.

(B) **Cardy's formula.** Smirnov's proof [56] of Cardy's formula has resisted extension to models beyond site percolation on the triangular lattice. It seems likely that Cardy's formula is valid for canonical percolation on any reasonable isoradial graph. There is a strong sense in which the existence of interfaces is preserved under the star–triangle transformations of the proofs. On the other hand, there is currently only limited control of the geometrical perturbations of interfaces, and in addition Cardy's formula is as yet unproven for *all* isoradial bond percolation models.

(C) **The bounded-angles property.** It is normal in working with probability and isoradial graphs to assume the BAP, see for example [15]. In the language of finite element methods, [9], the BAP is an example of the Ženíšek–Zlámal condition.

The BAP is a type of uniform non-flatness assumption. It implies an equivalence of metrics, and enables a uniform boundedness of certain probabilities. It may, however, not be *necessary* for the box-crossing property, and hence for the main results above.

As a test case, consider the situation in which all rhombi have angles exactly ϵ and $\pi - \epsilon$. In the limit as $\epsilon \downarrow 0$, we obtain[2] the critical space–time percolation process on $\mathbb{Z} \times \mathbb{R}$, see Figure 5.1 and, for example, [24]. Let $B_n(\alpha)$ be an $n \times n$ square of $\mathbb{R}^2$ inclined at angle α, and let $C_n(\alpha)$ be the event that the square is traversed by an open path between two given opposite faces. It is elementary using duality that

$$\mathbb{P}\big(C_n(\tfrac{1}{4}\pi)\big) \to \tfrac{1}{2} \qquad \text{as } n \to \infty.$$

Numerical simulations (of A. Holroyd) suggest that the same limit holds when $\alpha = 0$. A proof of this would suggest that the limit does not depend on α, and this in turn would support the possibility that the critical space–time percolation process satisfies Cardy's formula.

(D) **The square-grid property.** The SGP is a useful tool in the proof of Theorem 4.2, but it may not be necessary. In [29] is presented an isoradial graph without the SGP, and this example may be handled using an additional ad hoc argument.

[2] Joint work with Omer Angel.

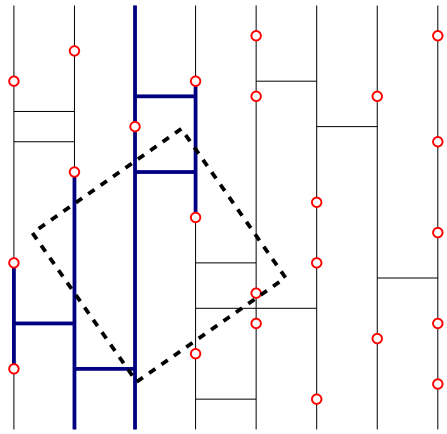


Figure 5.1. *Space–time percolation.* Each line is cut at rate 1, and nearest neighbours are joined at rate 1. One of the open clusters is highlighted. We ask for the probability that the box is traversed by an open path from its lower left side to its upper right side.

(E) **Near-critical exponents.** Theorem 4.2 establishes the universality of exponents *at criticality*. The method of proof does not appear to be extendable to the *near-critical exponents*, and it is an open problem to prove these to be universal for isoradial graphs. Kesten showed in [40] (see also [47]) that certain properties of a critical percolation process imply properties of the near-critical process, so long as the underlying graph has a sufficiently rich automorphism group. In particular, for such graphs, knowledge of certain critical exponents at criticality implies knowledge of exponents away from criticality. Only certain special isoradial graphs have sufficient homogeneity for such arguments to hold without new ideas of substance, and it is an open problem to weaken these assumptions of homogeneity. See the discussion around [28, Thm 1.2].

(F) **Random-cluster models.** How far may the proofs be extended to other models? It may seem at first sight that only a star–triangle transformation is required, but, as usual in such situations, boundary conditions play a significant role for dependent models such as the random-cluster model. The control of boundary conditions presents a new difficulty, so far unexplained. We return to this issue in Section 7.

6. Random-cluster model

6.1. Background. The random-cluster model was introduced by Fortuin and Kasteleyn around 1970 as a unification of processes satisfying versions of the series and parallel laws. In its base form, the random-cluster model has two parameters, an edge-parameter p and a cluster-weighting factor q.

Let $G = (V, E)$ be a finite graph, with associated configuration space $\Omega = \{0, 1\}^E$. For $\omega \in \Omega$ and $e \in E$, the edge e is designated *open* if $\omega_e = 1$. Let $k(\omega)$ be the number of open clusters of a configuration ω. The *random-cluster measure* on Ω, with parameters $p \in [0, 1]$,

$q \in (0, \infty)$, is the probability measure satisfying

$$\phi_{p,q}(\omega) \propto q^{k(\omega)} \mathbb{P}_p(\omega), \qquad \omega \in \Omega, \tag{6.1}$$

where $\mathbb{P}_p$ is the percolation product-measure with density p. In a more general setting, each edge $e \in E$ has an associated parameter p_e.

Bond percolation is retrieved by setting $q = 1$, and electrical networks arise via the limit $p, q \to 0$ in such a way that $q/p \to 0$. The relationship to Ising/Potts models is more complicated and involves a transformation of measures. In brief, two-point connection probabilities for the random-cluster measure with $q \in \{2, 3, \dots\}$ correspond to two-point correlations for ferromagnetic q-state Ising/Potts models, and this allows a geometrical interpretation of the latter's correlation structure. A fuller account of the random-cluster model and its history and associations may be found in [23, 64], to which the reader is referred for the basic properties of the model.

The special cases of percolation and the Ising model are very much better understood than is the general random-cluster model. We restrict ourselves to two-dimensional systems in this review, and we concentrate on the question of the identification of critical surfaces for certain isoradial graphs.

Two pieces of significant recent progress are reported here. Firstly, Beffara and Duminil-Copin [4] have developed the classical approach of percolation in order to identify the critical point of the square lattice, thereby solving a longstanding conjecture. Secondly, together with Smirnov [5], they have made use of the so-called parafermionic observable of [58] in a study of the critical surfaces of periodic isoradial graphs with $q \geq 4$.

6.2. Formalities. The random-cluster measure may not be defined directly on an *infinite* graph G. There are two possible ways to proceed in the setting of an infinite graph, namely via either boundary conditions or the DLR condition. The former approach works as follows. Let $(G_n : n \geq 1)$ be an increasing family of finite subgraphs of G that exhaust G in the limit $n \to \infty$, and let ∂G_n be the *boundary* of G_n, that is, ∂G_n is the set of vertices of G_n that are adjacent to a vertex of G not in G_n. A *boundary condition* is an equivalence relation b_n on ∂G_n; any two vertices $u, v \in \partial G_n$ that are equivalent under b_n are taken to be part of the same cluster. The extremal boundary conditions are: the *free* boundary condition, denoted $b_n = 0$, for which each vertex is in a separate equivalence class; and the *wired* boundary condition, denoted $b_n = 1$, with a unique equivalence class. We now consider the set of weak limits as $n \to \infty$ of the random-cluster measures on G_n with boundary conditions b_n.

Assume henceforth that $q \geq 1$. Then the random-cluster measures have properties of positive association and stochastic ordering, and one may deduce that the free and wired boundary conditions $b_n = 0$ and $b_n = 1$ are extremal in the following sense: (i) there is a unique weak limit of the free measures (respectively, the wired measures), and (ii) any other weak limit lies, in the sense of stochastic ordering, between these two limits. We write $\phi^0_{p,q}$ and $\phi^1_{p,q}$ for the free and wired weak limits. It is an important question to determine when $\phi^0_{p,q} = \phi^1_{p,q}$, and the answer so far is incomplete even when G has a periodic structure, see [23, Sect. 5.3].

The *percolation probabilities* are defined by

$$\theta^b(p, q) = \phi^b_{p,q}(0 \leftrightarrow \infty), \qquad b = 0, 1, \tag{6.2}$$

and the *critical values* by

$$p^b_{\mathrm{c}}(q) = \sup\{p : \theta^b(p, q) = 0\}, \qquad b = 0, 1. \tag{6.3}$$

Suppose that G is embedded in $\mathbb{R}^d$ in a natural manner. When G is *periodic* (that is, its embedding is invariant under a $\mathbb{Z}^d$ action), there is a general argument using convexity of pressure (see [21]) that implies that $p_c^0(q) = p_c^1(q)$, and in this case we write $p_c(q)$ for the common value.

One of the principal problems is to determine for which q the percolation probability $\theta^1(p, q)$ is discontinuous at the critical value p_c. This amounts to asking when $\theta^1(p_c, q) > 0$; the phase transition is said to be of *first order* whenever the last inequality holds. The phase transition is known to be of first order for sufficiently large q, and is believed to be so if and only if $q > Q(d)$ for some $Q(d)$ depending on the dimension d. Furthermore, it is expected that

$$Q(d) = \begin{cases} 4 & \text{if } d = 2, \\ 2 & \text{if } d \geq 4. \end{cases}$$

We restrict our attention henceforth to the case $d = 2$, for which it is believed that the value $q = 4$ separates the first and second order transitions. Recall Conjecture 2.1 and note the recent proof that $Q(2) \geq 4$, for which the reader is referred to [18] and the references therein.

6.3. Critical point on the square lattice. The square lattice $\mathbb{Z}^2$ is one of the main playgrounds of physicists and probabilists. Although the critical points of percolation, the Ising model and some Potts models on $\mathbb{Z}^2$ are long proved, the general answer for random-cluster models (and hence all Potts models) has been proved only recently.

Theorem 6.1 (Criticality [4]). *The random-cluster model on the square lattice with cluster-weighting factor $q \geq 1$ has critical value*

$$p_c(q) = \frac{\sqrt{q}}{1 + \sqrt{q}}.$$

This exact value has been 'known' for a long time. When $q = 1$, the statement $p_c(1) = \frac{1}{2}$ is the Harris–Kesten theorem for bond percolation. When $q = 2$, it amounts to the well known calculation of the critical temperature of the Ising model. For large q, the result (and more) was proved in [42, 43] ($q > 25.72$ suffices, see [23, Sect. 6.4]). There is a 'physics proof' in [34] for $q \geq 4$.

The main contribution of [4] is a proof of a box-crossing property using a clever extension of the 'RSW' arguments of Russo and Seymour–Welsh in the context of the symmetry illustrated in Figure 6.1, combined with careful control of boundary conditions. An alternative approach is developed in [19].

6.4. Isoradiality and the star–triangle transformation. The star–triangle transformation for the random-cluster model is similar to that of percolation, and is illustrated in Figure 6.2. The three edges of the triangle have parameters p_0, p_1, p_2, and we set $\mathbf{y} = (y_0, y_1, y_2)$ where

$$y_i = \frac{p_i}{1 - p_i}.$$

The corresponding edges of the star have parameters y_i' where $y_i y_i' = q$. Finally, we require that the y_i satisfy $\psi(\mathbf{y}) = 0$ where

$$\psi(\mathbf{y}) = y_1 y_2 y_3 + y_1 y_2 + y_2 y_3 + y_3 y_1 - q. \tag{6.4}$$

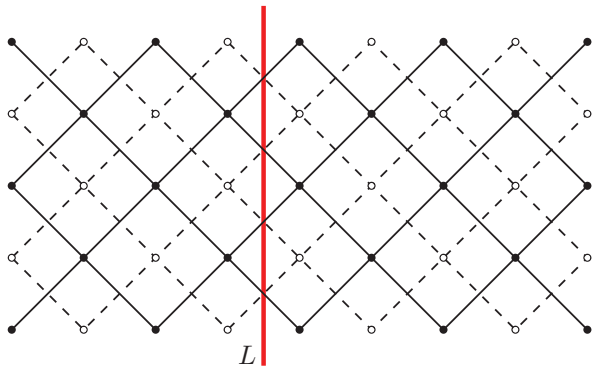


Figure 6.1. The square lattice and its dual, rotated through $\pi/4$. Under reflection in the line L, the primal is mapped to the dual.

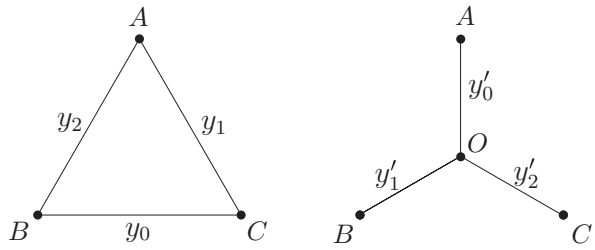


Figure 6.2. The star–triangle transformation for the random-cluster model.

Further details of the star–triangle transformation for the random-cluster model may be found in [23, Sect. 6.6].

We now follow the discussion of Section 4.2 of the relationship between the star–triangle transformation and the rhombus-switch of Figure 4.1. In so doing, we arrive (roughly as in [36, p. 282]) at the 'right' parametrization for an isoradial graph G, namely with (3.1) replaced by

$$\begin{aligned} \text{if } 1 \le q < 4: \quad & y_e = \sqrt{q}\,\frac{\sin(\frac{1}{2}\sigma(\pi-\theta_e))}{\sin(\frac{1}{2}\sigma\theta_e)}, \qquad \cos(\tfrac{1}{2}\sigma\pi) = \tfrac{1}{2}\sqrt{q}, \\ \text{if } q > 4: \quad & y_e = \sqrt{q}\,\frac{\sinh(\frac{1}{2}\sigma(\pi-\theta_e))}{\sinh(\frac{1}{2}\sigma\theta_e)}, \qquad \cosh(\tfrac{1}{2}\sigma\pi) = \tfrac{1}{2}\sqrt{q}, \end{aligned} \tag{6.5}$$

where θ_e is given in Figure 3.3. The intermediate case $q = 4$ is the common limit of the two expressions as $q \to 4$, namely

$$y_e = 2\,\frac{\pi - \theta_e}{\theta_e}.$$

Write $\phi^b_{G,q}$ for the corresponding random-cluster measure on an isoradial graph G with boundary condition $b = 0, 1$. We refer to $\phi^0_{G,q}$ as the 'canonical random-cluster measure' on G.

6.5. Criticality via the parafermion. Theorem 6.1 is proved in [4] by classical methods, and it holds for all $q \geq 1$. The proof is sensitive to the assumed symmetries of the lattice, and does not currently extend even to the inhomogeneous random-cluster model on $\mathbb{Z}^2$ in which the vertical and horizontal edges have different parameter values. In contrast, the parafermionic observable introduced by Smirnov [58] has been exploited by Beffara, Duminil-Copin, and Smirnov [5] to study the critical point of fairly general isoradial graphs subject to the condition $q \geq 4$.

Let $G = (V, E)$ be an isoradial graph. For $\beta \in (0, \infty)$, let $y_e(\beta) = \beta y_e$ where y_e is given in (6.5). Let

$$p_e(\beta) = \frac{y_e(\beta)}{1 + y_e(\beta)}$$

accordingly, and write $\phi^b_{G,q,\beta}$ for the corresponding random-cluster measure on G with boundary condition b. The following result of [5] is proved by a consideration of the parafermionic observable.

Theorem 6.2 ([5]). *Let $q \geq 4$, and let G be an isoradial graph satisfying the BAP. For $\beta < 1$, there exists $a > 0$ such that*

$$\phi^0_{G,q,\beta}(u \leftrightarrow v) \leq e^{-a|u-v|}, \qquad u, v \in V.$$

One deduces from Theorem 6.2 using duality that

(a) for $\beta < 1$, $\phi^0_{G,q,\beta}$-a.s., there is no infinite open cluster, and

(b) for $\beta > 1$, $\phi^1_{G,q,\beta}$-a.s., there exists a unique infinite open cluster.

This is only a partial verification of the criticality of the canonical measure, since parts (a) and (b) deal with potentially different measures, namely the free and wired limit measures, respectively. Further progress may be made for *periodic* graphs, as follows. Subject to the assumption of periodicity, it may be proved as in [21] that $\phi^0_{G,q,\beta} = \phi^1_{G,q,\beta}$ for almost every β, and hence that part (b) holds with $\phi^1_{G,q,\beta}$ replaced by $\phi^0_{G,q,\beta}$. Therefore, for periodic embeddings, the canonical measure $\phi^0_{G,q} = \phi^0_{G,q,1}$ is critical.

Here is an application of the above remarks to the (periodic) inhomogeneous square lattice.

Corollary 6.3 ([5]). *Let $q \geq 4$, and consider the random-cluster model on $\mathbb{Z}^2$ with the variation that horizontal edges have parameter p_1 and vertical edges parameter p_2. The critical surface is given by $y_1 y_2 = q$ where $y_i = p_i/(1 - p_i)$.*

We close with the observation that a great deal more is known in the special case when $q = 2$. The $q = 2$ random-cluster model corresponds to the Ising model, for which the special arithmetic of the equation $1 + 1 = 2$ permits a number of techniques which are not available in greater generality. In particular, the Ising model and the $q = 2$ random-cluster model on an isoradial graph lend themselves to a fairly complete theory using the parafermionic observable. The interested reader is directed to the work of Smirnov [57, 58] and Chelkak–Smirnov [16].

7. Open problems for the random-cluster model

(A) **Inhomogeneous models.** Extend Corollary 6.3 to cover the case $1 \leq q < 4$.

(B) **Periodicity.** Remove the assumption of periodicity in the proof of criticality of the canonical random-cluster measure on isoradial graphs. It would suffice to prove that $\phi^0_{G,q,\beta} = \phi^1_{G,q,\beta}$ for almost every β, without the assumption of periodicity. More generally, it would be useful to have a proof of the uniqueness of Gibbs states for *aperiodic* interacting systems, along the lines of that of Lebowitz and Martin-Löf [46] for a *periodic* Ising model.

(C) **Bounded-angles property.** Remove the assumption of the bounded-angles property in Theorem 6.1.

(D) **Criticality and universality for general q.** Adapt the arguments of [29] (or otherwise) to prove criticality and universality for the canonical random-cluster measure on isoradial graphs either for general $q \geq 1$ or subject to the restriction $q \geq 4$.

Acknowledgements. The author is grateful to Ioan Manolescu for many discussions concerning percolation on isoradial graphs, and to Omer Angel and Alexander Holroyd for discussions about the space–time percolation process of Figure 5.1. Hugo Duminil-Copin and Ioan Manolescu kindly commented on a draft of this paper. This work was supported in part by the EPSRC under grant EP/103372X/1.

References

[1] R. Bauerschmidt, H. Duminil-Copin, J. Goodman, and G. Slade. Lectures on self-avoiding-walks. In D. Ellwood, C. M. Newman, V. Sidoravicius, and W. Werner, editors, *Probability and Statistical Physics in Two and More Dimensions*, volume 15 of Clay Mathematics Institute Proceedings, pp. 395–476. CMI/AMS publication, 2012.

[2] R. J. Baxter, *Solvable eight-vertex model on an arbitrary planar lattice*, Philos. Trans. Roy. Soc. London Ser. A **289** (1978), 315–346.

[3] ———, *Exactly Solved Models in Statistical Mechanics*, Academic Press, London, 1982.

[4] V. Beffara and H. Duminil-Copin, *The self-dual point of the two-dimensional random-cluster model is critical for $q \geq 1$*, Probab. Th. Rel. Fields, **153** (2012), 511–542.

[5] V. Beffara, H. Duminil-Copin, and S. Smirnov, *On the critical parameters of the $q \geq 4$ random-cluster model on isoradial graphs*, 2013, preprint.

[6] B. Bollobás and O. Riordan, *Percolation*, Cambridge University Press, Cambridge, 2006.

[7] C. Boutillier and B. de Tilière, *The critical Z-invariant Ising model via dimers: Locality property*, Commun. Math. Phys. **301** (2011), 473–516.

[8] C. Boutillier and B. de Tilière, *Statistical mechanics on isoradial graphs*, In J.-D. Deuschel, B. Gentz, W. König, M. von Renesse, M. Scheutzow, and U. Schmock, editors, Probability in Complex Physical Systems, volume 11 of Springer Proceedings in Mathematics, pp. 491–512, 2012.

[9] J. Brandts, S. Korotov, and M. Křížek, *Generalization of the Zlámal condition for simplicial finite elements in $\mathbb{R}^d$*, Applic. Math. **56** (2011), 417–424.

[10] S. R. Broadbent and J. M. Hammersley, *Percolation processes I. Crystals and mazes*,

Proc. Camb. Phil. Soc. **53** (1957), 629–641.

[11] N. G. de Bruijn, *Algebraic theory of Penrose's non-periodic tilings of the plane. I*, Indagat. Math. (Proc.) **84** (1981), 39–52.

[12] ———, *Algebraic theory of Penrose's non-periodic tilings of the plane. II*, Indagat. Math. (Proc.) **84** (1981), 53–66.

[13] ———, *Dualization of multigrids*, J. Phys. Colloq. **47** (1986), 85–94.

[14] J. Cardy, *Critical percolation in finite geometries*, J. Phys. A: Math. Gen. **25** (1992), L201–L206.

[15] D. Chelkak and S. Smirnov, *Discrete complex analysis on isoradial graphs*, Adv. Math. **228** (2011), 1590–1630.

[16] ———, *Universality in the 2D Ising model and conformal invariance of fermionic observables*, Invent. Math. **189** (2012), 515–580.

[17] R. J. Duffin, *Potential theory on a rhombic lattice*, J. Combin. Th. **5** (1968), 258–272.

[18] H. Duminil-Copin, *Parafermionic observables and their applications to planar statistical physics models*, Ensaios Matemáticos **25** (2013), 1–371.

[19] H. Duminil-Copin and I. Manolescu, *The phase transition of the planar random-cluster model and Potts model with $q \geq 1$ is sharp*, 2014, in preparation.

[20] R. J. Fitzner, *Non-backtracking lace expansion*, PhD thesis, Technische Universiteit Eindhoven, 2013.

[21] G. R. Grimmett, *The stochastic random-cluster process and the uniqueness of random-cluster measures*, Ann. Probab. **23** (1995), 1461–1510.

[22] ———, *Percolation*, Springer, Berlin, 2nd edition, 1999.

[23] ———, *The Random-Cluster Model*, Springer, Berlin, 2006.

[24] ———, *Space–time percolation*, In V. Sidoravicius and M. E. Vares, editors, In and Out of Equilibrium 2, volume 60 of Progress in Probability, pp. 305–320, Birkhäuser, Boston, 2008.

[25] ———, *Probability on Graphs*, Cambridge University Press, Cambridge, 2010. http://www.statslab.cam.ac.uk/~grg/books/pgs.html.

[26] ———, *Three theorems in discrete random geometry*, Probab. Surveys **8** (2011), 403–441.

[27] G. R. Grimmett and I. Manolescu, *Inhomogeneous bond percolation on the square, triangular, and hexagonal lattices*, Ann. Probab. **41** (2013), 2990–3025.

[28] ———, *Universality for bond percolation in two dimensions*, Ann. Probab. **41** (2013), 3261–3283.

[29] ———, *Bond percolation on isoradial graphs: criticality and universality*, Probab. Th. Rel. Fields **159** (2014), 273–327.

[30] J. M. Hammersley, *Percolation processes. Lower bounds for the critical probability*, Ann. Math. Statist. **28** (1957), 790–795.

[31] ———, *Bornes supérieures de la probabilité critique dans un processus de filtration*, In Le Calcul des Probabilités et ses Applications, pp. 17–37, CNRS, Paris, 1959.

[32] T. Hara and G. Slade, *Mean-field critical behaviour for percolation in high dimensions*, Commun. Math. Phys. **128** (1990), 333–391.

[33] ———, *Mean-field behaviour and the lace expansion*, In G. R. Grimmett, editor, Probability and Phase Transition, pp. 87–122, Kluwer, 1994.

[34] D. Hintermann, H. Kunz, and F. Y. Wu, *Exact results for the Potts model in two dimensions*, J. Statist. Phys. **19** (1978), 623–632.

[35] A. E. Kennelly, *The equivalence of triangles and three-pointed stars in conducting networks*, Electrical World and Engineer **34** (1899), 413–414.

[36] R. Kenyon, *An introduction to the dimer model*, In G. F. Lawler, editor, School and Conference on Probability Theory, volume 17 of Lecture Notes Series, pp. 268–304. ICTP, Trieste, 2004. http://publications.ictp.it/lns/vol17/vol17toc.html.

[37] R. Kenyon and J.-M. Schlenker, *Rhombic embeddings of planar quad-graphs*, Trans. Amer. Math. Soc. **357** (2005), 3443–3458.

[38] H. Kesten, *The critical probability of bond percolation on the square lattice equals* $1/2$, Commun. Math. Phys. **74** (1980), 44–59.

[39] ———, *Percolation Theory for Mathematicians*, Birkhäuser, Boston, 1982.

[40] ———, *Scaling relations for 2D-percolation*, Commun. Math. Phys. **109** (1987), 109–156.

[41] G. Kozma and A. Nachmias, *Arm exponents in high dimensional percolation*, J. Amer. Math. Soc. **24** (2011), 375–409.

[42] L. Laanait, A. Messager, S. Miracle-Solé, J. Ruiz, and S. Shlosman, *Interfaces in the Potts model I: Pirogov–Sinai theory of the Fortuin–Kasteleyn representation*, Commun. Math. Phys. **140** (1991), 81–91.

[43] L. Laanait, A. Messager, and J. Ruiz, *Phase coexistence and surface tensions for the Potts model*, Commun. Math. Phys. **105** (1986), 527–545.

[44] B. Laslier and F. B. Toninelli, *Lozenge tilings, Glauber dynamics and macroscopic shape*, 2013. arXiv:1310.5844.

[45] G. F. Lawler, O. Schramm, and W. Werner, *One-arm exponent for 2D critical percolation*, Electron. J. Probab. 7:Paper 2, 2002.

[46] J. L. Lebowitz and A. Martin-Löf, *On the uniqueness of the equilibrium state for Ising spin systems*, Commun. Math. Phys. **25** (1972), 276–282.

[47] P. Nolin, *Near-critical percolation in two dimensions*, Electron. J. Probab. **13** (2008), 1562–1623.

[48] L. Onsager, *Crystal statistics. I. A two-dimensional model with an order–disorder transition*, Phys. Rev. **65** (1944), 117–149.

[49] R. Penrose, *The rôle of aesthetics in pure and applied mathematical research*, Bull. Inst. Math. Appl. **10** (1974), 266–271.

[50] ———, *Pentaplexity*, Eureka **39** (1978), 16–32, reprinted in Math. Intellig. 2 (1979), 32–37.

[51] J. H. H. Perk and H. Au-Yang, *Yang–Baxter equation*, In J.-P. Françoise, G. L. Naber, and S. T. Tsou, editors, Encyclopedia of Mathematical Physics, volume 5, pp. 465–473. Elsevier, 2006.

[52] L. Russo, *A note on percolation*, Z. Wahrsch'theorie verw. Geb. **43** (1978), 39–48.

[53] O. Schramm, *Scaling limits of loop-erased walks and uniform spanning trees*, Israel J.

Math. **118** (2000), 221–288.

[54] ———, *Conformally invariant scaling limits: an overview and collection of open problems*, In M. Sanz-Solé et al., editor, Proceedings of the International Congress of Mathematicians, Madrid, volume I, pp. 513–544. European Mathematical Society, Zurich, 2007.

[55] P. D. Seymour and D. J. A. Welsh, *Percolation probabilities on the square lattice*, Ann. Discrete Math. **3** (1978), 227–245.

[56] S. Smirnov, *Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits*, C. R. Acad. Sci. Paris Ser. I Math. **333** (2001), 239–244.

[57] ———, *Towards conformal invariance of 2D lattice models*, In M. Sanz-Solé et al., editor, Proceedings of the International Congress of Mathematicians, Madrid, 2006, volume II, pp. 1421–1452. European Mathematical Society, Zurich, 2007.

[58] ———, *Conformal invariance in random cluster models. I. Holomorphic fermions in the Ising model*, Ann. Math. **172** (2010), 1435–1467.

[59] ———, *Discrete complex analysis and probability*, In R. Bhatia, A. Pal, G. Rangarajan, V. Srinivas, and M. Vanninathan, editors, Proceedings of the International Congress of Mathematicians, Hyderabad, 2010, volume I, pp. 595–621. Hindustan Book Agency, New Delhi, 2010.

[60] S. Smirnov and W. Werner, *Critical exponents for two-dimensional percolation*, Math. Res. Lett. **8** (2001), 729–744.

[61] N. Sun, *Conformally invariant scaling limits in planar critical percolation*, Probab. Surveys, **8** (2011), 155–209.

[62] M. F. Sykes and J. W. Essam, *Some exact critical percolation probabilities for site and bond problems in two dimensions*, J. Math. Phys. **5** (1964), 1117–1127.

[63] W. Werner, *Lectures on two-dimensional critical percolation*, In S. Sheffield and T. Spencer, editors, Statistical Mechanics, volume 16, pp. 297–360. IAS–Park City, 2007.

[64] ———, *Percolation et Modèle d'Ising*, volume 16 of Cours Specialisés, Société Mathématique de France, Paris, 2009.

[65] R. M. Ziff, C. R. Scullard, J. C. Wierman, and M. R. A. Sedlock, *The critical manifolds of inhomogeneous bond percolation on bow-tie and checkerboard lattices*, J. Phys. A **45** (2012), 494005.

Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK
E-mail: g.r.grimmett@statslab.cam.ac.uk

Singular stochastic PDEs

Martin Hairer

Abstract. We present a series of recent results on the well-posedness of very singular parabolic stochastic partial differential equations. These equations are such that the question of what it even means to be a solution is highly non-trivial. This problem can be addressed within the framework of the recently developed theory of "regularity structures", which allows to describe candidate solutions locally by a "jet", but where the usual Taylor polynomials are replaced by a sequence of custom-built objects. In order to illustrate the theory, we focus on the particular example of the Kardar-Parisi-Zhang equation, a popular model for interface propagation.

Mathematics Subject Classification (2010). 60H15, 81S20, 82C28.

Keywords. Regularity structures, renormalisation, stochastic PDEs.

1. Introduction

In this article, we report on a recently developed theory [23] allowing to give a robust meaning to a large class of stochastic partial differential equations (SPDEs) that have traditionally been considered to be ill-posed. The general structure of these equations is

$$\mathcal{L}u = F(u) + G(u)\xi \,, \tag{1.1}$$

where the dominant linear operator $\mathcal{L}$ is of parabolic (or possibly elliptic) type, F and G are local nonlinearities depending on u and its derivatives of sufficiently low order, and ξ is some driving noise. Problems arise when ξ (and therefore also u) is so singular that some of the terms appearing in F and / or the product between G and ξ are ill-posed. For simplicity, we will consider all of our equations in a *bounded* spatial region with periodic boundary conditions.

One relatively simple example of an ill-posed equation of the type (1.1) is that of a system of equations with a nonlinearity of Burgers type driven by space-time white noise:

$$\partial_t u = \partial_x^2 u + F(u)\,\partial_x u + \xi \,. \tag{1.2}$$

(See Section 2.2 below for a definition of the space-time white noise ξ.) Here, $u(x,t) \in \mathbf{R}^n$ and F is a smooth matrix-valued function, so that one can in general not rewrite the nonlinearity as a total derivative. In this example, which was originally studied in [20] but then further analysed in the series of articles [24, 25, 29], solutions at any fixed instant of time have exactly the same regularity (in space) as Brownian motion. As a consequence, $\partial_x u$ is expected to "look like" white noise. It is of course very well-known from the study of

Proceedings of the International Congress of Mathematicians, Seoul, 2014

ordinary stochastic differential equations (SDEs) that in this case the product $F(u)\,\partial_x u$ is "unstable": one can get different answers depending on the type of limiting procedure used to define it. This is the reason why one has different solution theories for SDEs: one obtains different answers, depending on whether they are interpreted in the Itô or in the Stratonovich sense [30, 43, 44].

Another example is given by the KPZ equation [32] which can formally be written as

$$\partial_t h = \partial_x^2 h + (\partial_x h)^2 - C + \xi \,, \tag{1.3}$$

and is a very popular model of one-dimensional interface propagation. As in the case of (1.2), one expects solutions to this equation to "look like" Brownian motion (in space) for any fixed instant of time. Now the situation is much worse however: the nonlinearity looks like the square of white noise, which really shouldn't make any sense! In this particular case however, one can use a "trick", the Cole-Hopf transform, to reduce the problem to an equation that has an interpretation within the framework of classical SPDE theory [4]. Furthermore, this "Cole-Hopf solution" was shown in [4] to be the physically relevant solution since it describes the mesoscopic fluctuations of a certain microscopic interface growth model, see also [17]. On the other hand, the problem of interpreting these solutions directly at the level of (1.3) and to show their stability under suitable approximations had been open for a long time, before being addressed in [21].

Both examples mentioned so far have only one space dimension. This particular feature (together with some additional structure in the case of the KPZ equation, see Remark 5.17 below) allowed to treat them by borrowing estimates and techniques from the theory of (controlled) rough paths [15, 18, 34]. This approach breaks down in higher spatial dimensions. More recently, a general theory of "regularity structures" was developed in [23], which unifies many previous approaches and allows in particular to treat higher dimensional problems.

Two nice examples of equations that can be treated with this new approach are given by

$$\partial_t \Phi = \Delta \Phi + C\Phi - \Phi^3 + \xi \,, \tag{1.4a}$$

$$\partial_t \Psi = -\Delta\big(\Delta \Psi + C\Psi - \Psi^3\big) + \operatorname{div} \xi \,, \tag{1.4b}$$

in space dimension $d = 3$. These equations can be interpreted as the natural "Glauber" and "Kawasaki" dynamics associated to Euclidean Φ^4 field theory in the context of stochastic quantisation [40]. It is also expected to describe the dynamical mesoscale fluctuations for phase coexistence models that are "almost mean-field", see [5]. These equations cease to have function-valued solutions in dimension $d \geq 2$, so that the classical interpretation of the cubic nonlinearity loses its meaning there. In two dimensions, a solution theory for these equations was developed in [1], which was later improved in [10–12], see Section 3.1 below. The case $d = 3$ (which is the physically relevant one in the interpretation as dynamical fluctuations for phase coexistence models) had remained open and was eventually addressed in [23].

A final example of the kind of equations that can be addressed by the theory exposed in these notes (but this list is of course not exhaustive) is a continuous analogue to the classical parabolic Anderson model [8]:

$$\partial_t u = \Delta u + u\,\eta + Cu \,, \tag{1.5}$$

in dimensions $d \in \{2, 3\}$. In this equation, η denotes a noise term that is white in space, but constant in time. This time, the problem is that in dimension $d \geq 2$, the product $u\,\eta$ ceases to make sense classically, as a consequence of the lack of regularity of u.

The following "meta-theorem" (formulated in a somewhat vague sense, precise formulations differ slightly from problem to problem and can be found in the abovementioned articles) shows in which sense one can give meaning to all of these equations.

Theorem 1.1. *Consider the sequence of classical solutions to any of the equations* (1.2)–(1.5) *with ξ (resp. η) replaced by a smooth regularised noise ξ_ε and $C = C_\varepsilon$ depending on ε. Then, there exists a choice $C_\varepsilon \to \infty$ such that this sequence of solutions converges to a limit in probability, locally in time. Furthermore, this limit is universal, i.e. does not depend on the details of the regularisation ξ_ε.*

Besides these convergence results, the important fact here is that the limit is *independent* of the precise details of the regularisation mechanism. In addition, the theory of regularity structures also yields rates of convergence, as well as an intrinsic description of these limits. It also provides automatically a very detailed local description of these limits.

The aim of this article is to give an overview of the ingredients involved in the proof of a result like Theorem 1.1. We structure this as follows. In Section 2, we recall a number of properties and definitions of Hölder spaces of positive (and negative!) order that will be useful for our argument. In Section 3, we then explain how, using only standard tools, it is possible to provide a robust solution theory for not-so-singular SPDEs, like for example (1.4) in dimension $d = 2$. Section 4 is devoted to a short overview of the main definitions and concepts of the abstract theory of regularity structures which is a completely general way of formalising the properties of objects that behave "like Taylor polynomials". Section 5 then finally shows how one can apply this general theory to the specific context of the type of parabolic SPDEs considered above, how renormalisation procedures can be built into the theory, and how this affects the equations.

Throughout the whole article, our argumentation will remain mostly at the heuristic level, but we will make the statements and definitions as precise as possible.

1.1. An alternative approach. A different approach to building solution theories for singular PDEs was developed simultaneously to the one presented here by Gubinelli & Al in [19]. That approach is based on the properties of Bony's paraproduct [2, 3, 7], in particular on the paralinearisation formula. One advantage is that in the paraproduct-based approach one generally deals with globally defined objects rather than the "jets" used in the theory of regularity structures. This comes at the expense of achieving a less clean break between the analytical and the algebraic aspects of a given problem and obtaining less detailed information about the solutions. Furthermore, its scope is not as wide as that of the theory of regularity structures, see also Remark 5.17 below for more details.

2. Some properties of Hölder spaces

We recall in this section a few standard results from harmonic analysis that are very useful to have in mind. Note first that the linear part of all of the equations described in the introduction is invariant under some space-time scaling. In the case of the heat equation, this is the parabolic scaling. In other words, if u is a solution to the heat equation, then $\tilde{u}(t,x) = u(\lambda^{-2}t, \lambda^{-1}x)$ is also a solution to the heat equation.

This suggests that we should look for solutions in function / distribution spaces respecting this scaling. Given a smooth compactly supported test function φ and a space-time

coordinate $z = (t, x)$, we henceforth write $\varphi_z^\lambda(s, y) = \lambda^{-d-2}\varphi\big(\lambda^{-2}(s-t), \lambda^{-1}(y-x)\big)$, where d denotes the spatial dimension and the factor λ^{-d-2} is chosen so that the integral of φ_z^λ is the same as that of φ. In the case of the stochastic Cahn-Hilliard equation (1.4b), we would naturally use instead a temporal scaling of λ^{-4} and the prefactor would then be λ^{-d-4}.

With these notations at hand, we define spaces of distributions $\mathcal{C}^\alpha$ for $\alpha < 0$ in the following way. Denoting by $\mathcal{B}_\alpha$ the set of smooth test functions $\varphi\colon \mathbf{R}^{d+1} \to \mathbf{R}$ that are supported in the centred ball of radius 1 and such that their derivatives of order up to $1 + |\alpha|$ are uniformly bounded by 1, we set

Definition 2.1. Let η be a distribution on $d+1$-dimensional space-time and let $\alpha < 0$. We say that $\eta \in \mathcal{C}^\alpha$ if the bound $\big|\eta(\varphi_z^\lambda)\big| \lesssim \lambda^\alpha$ holds uniformly over all $\lambda \in (0, 1]$, all $\varphi \in \mathcal{B}_\alpha$, and locally uniformly over $z \in \mathbf{R}^{d+1}$.

For $\alpha \geq 0$, we say that a function $f\colon \mathbf{R}^{d+1} \to \mathbf{R}$ belongs to $\mathcal{C}^\alpha$ if, for every $z \in \mathbf{R}^{d+1}$ there exists a polynomial P_z of (parabolic) degree at most α and such that the bound

$$|f(z') - P_z(z')| \lesssim |z - z'|^\alpha \,,$$

holds locally uniformly over z and uniformly over all z' with $|z' - z| \leq 1$. Here, we say that a polynomial P in $z = (t, x)$ is of parabolic degree n if each monomial is of the form z^k with $|k| = 2|k_0| + \sum_{i \neq 0} k_i \leq n$. In other words, the degree of the time variable "counts double". For $z = (t, x)$, we furthermore write $|z| = |t|^{1/2} + |x|$. (When treating (1.4b), powers of t count four times and one writes $|z| = |t|^{1/4} + |x|$.)

We now collect a few important properties of the spaces $\mathcal{C}^\alpha$.

2.1. Analytical properties. First, given a function and a distribution (or two distributions) it is natural to ask under what regularity assumptions one can give an unambiguous meaning to their product. It is well-known, at least in the Euclidean case but the extension to the parabolic case is straightforward, that the following result yields a sharp criterion for when, in the absence of any other structural knowledge, one can multiply a function and distribution of prescribed regularity [2, Thm 2.52].

Theorem 2.2. *Let $\alpha, \beta \neq 0$. Then, the map $(f, g) \mapsto f \cdot g$ defined on all pairs of continuous functions extends to a continuous bilinear map from $\mathcal{C}^\alpha \times \mathcal{C}^\beta$ to the space of all distributions if and only if $\alpha + \beta > 0$. Furthermore, if $\alpha + \beta > 0$, the image of the multiplication operator is $\mathcal{C}^{\alpha \wedge \beta}$.*

Another important property of these spaces is given by how they transform under convolution with singular kernels. Let $K\colon \mathbf{R}^{d+1} \to \mathbf{R}$ be a function that is smooth away from the origin and supported in the centred ball of radius 1. One should think of K as being a truncation of the heat kernel $\mathcal{G}$ in the sense that $\mathcal{G} = K + R$ where R is a smooth space-time function. We then say that K is of order β (in the case of a truncation of the heat kernel one has $\beta = 2$) if one can write $K = \sum_{n \geq 0} K_n$ for kernels K_n which are supported in the centred ball of radius 2^{-n} and such that

$$\sup_z |D^k K_n(z)| \lesssim 2^{((d+2)+|k|-\beta)n} \,, \tag{2.1}$$

for any fixed multiindex k, uniformly in n. Multiplying the heat kernel with a suitable partition of the identity, it is straightforward to verify that this bound is indeed satisfied.

With these notations at hand, one has the following very general Schauder estimate, see for example [41, 42] for special cases.

Theorem 2.3. *Let $\beta > 0$, let K be a kernel of order β, and let $\alpha \in \mathbf{R}$ be such that $\alpha+\beta \notin \mathbf{N}$. Then, the convolution operator $\eta \mapsto K \star \eta$ is continuous from $\mathcal{C}^\alpha$ into $\mathcal{C}^{\alpha+\beta}$.*

Remark 2.4. The condition $\alpha + \beta \notin \mathbf{N}$ seems somewhat artificial. It can actually be dispensed with by slightly changing the definition of $\mathcal{C}^\alpha$.

2.2. Probabilistic properties. Let now η be a random distribution, which we define in general as a continuous linear map $\varphi \mapsto \eta(\varphi)$ from the space of compactly supported smooth test functions into the space of square integrable random variables on some fixed probability space $(\Omega, \mathbf{P})$. We say that it satisfies *equivalence of moments* if, for every $p \geq 1$ there exists a constant C_p such that the bound

$$\mathbf{E}|\eta(\varphi)|^{2p} \leq C_p \big(\mathbf{E}|\eta(\varphi)|^2\big)^p \,,$$

holds for uniformly over all test functions φ. This is of course the case if the random variables $\eta(\varphi)$ are Gaussian, but it also holds if they take values in an inhomogeneous Wiener chaos of fixed order [39].

Given a stationary random distribution η and a (deterministic) distribution C, we say that η has covariance C if $\mathbf{E}\eta(\varphi)\eta(\psi) = \langle C \star \varphi, \psi\rangle$, where $\langle \cdot, \cdot \rangle$ denotes the L^2-scalar product. With this notation at hand, space-time white noise ξ is the Gaussian random distribution on $\mathbf{R}^{d+1}$ with covariance given by the delta distribution. In other words, $\xi(\varphi)$ is centred Gaussian for every φ and $\mathbf{E}\xi(\varphi)\xi(\psi) = \langle \varphi, \psi \rangle_{L^2}$.

Similarly to the case of stochastic processes, a random distribution $\tilde\eta$ is said to be a *version* of η if, for every fixed test function φ, $\tilde\eta(\varphi) = \eta(\varphi)$ almost surely. One then has the following Kolmogorov criterion, a proof of which can be found for example in [23].

Theorem 2.5. *Let η be a stationary random distribution satisfying equivalence of moments and such that, for some $\alpha < 0$, the bound*

$$\mathbf{E}|\eta(\varphi_z^\lambda)|^2 \lesssim \lambda^{2\alpha} \,,$$

holds uniformly over $\lambda \in (0, 1]$ and $\varphi \in \mathcal{B}_\alpha$. Then, for any $\kappa > 0$, there exists a $\mathcal{C}^{\alpha-\kappa}$-valued random variable $\tilde\eta$ which is a version of η.

From now on, we will make the usual abuse of terminology and not distinguish between different versions of a random distribution.

Remark 2.6. It follows immediately from the scaling properties of the L^2 norm that one can realise space-time white noise as a random variable in $\mathcal{C}^{-\frac{d}{2}-1-\kappa}$ for every $\kappa > 0$. This is sharp in the sense that it can *not* be realised as a random variable in $\mathcal{C}^{-\frac{d}{2}-1}$. This is akin to the fact that Brownian motion has sample paths belonging to $\mathcal{C}^\alpha$ for every $\alpha < \frac{1}{2}$, but *not* for $\alpha = \frac{1}{2}$.

Let now K be a kernel of order β as before, let ξ be space-time white noise, and set $\eta = K \star \xi$. It then follows from either Theorem 2.5 directly, or from Theorem 2.3 combined with Remark 2.6, that η belongs almost surely to $\mathcal{C}^\alpha$ for every $\alpha < \beta - \frac{d}{2} - 1$. We now turn to the question of how to define powers of η. If $\beta \leq \frac{d}{2} + 1$, η is not a random function, so that its powers are in general undefined.

Recall that if ξ is space-time white noise and $L^2(\xi)$ denotes the space of square-integrable random variables that are measurable with respect to the σ-algebra generated by ξ, then $L^2(\xi)$ can be decomposed into a direct sum $L^2(\xi) = \bigoplus_{m\geq 0} \mathcal{H}^m(\xi)$ so that $\mathcal{H}^0$ contains constants, $\mathcal{H}^1$ contains random variables of the form $\xi(\varphi)$ with $\varphi \in L^2$, and $\mathcal{H}^m$ contains suitable generalised Hermite polynomials of order m in the elements of $\mathcal{H}^1$, see [37, 39] for details. Elements of $\mathcal{H}^m$ have a representation by square-integrable kernels of m variables, and this representation is unique if we impose that the kernel is symmetric under permutation of its arguments. In other words, one has a surjection $I^{(m)}\colon L^2(\mathbf{R}^{d+1})^{\otimes m} \to \mathcal{H}^m$ and $I^{(m)}(L) = I^{(m)}(L')$ if and only if the symmetrisations of L and L' coincide.

In the particular case where K is non-singular, η is a random function and its nth power η^n can be represented as

$$\eta^n(\varphi) = \sum_{2m<n} P_{m,n} C^m\, I^{(n-2m)}(K_\varphi^{(n-2m)})\;, \tag{2.2}$$

where

$$K_\varphi^{(r)}(z_1,\dots,z_r) := \int K(z-z_1)\cdots K(z-z_r)\,\varphi(z)\,dz\;,$$

for some combinatorial factors $P_{m,n}$. Here we have set $C = \int K^2(z)\,dz$. A simple calculation then shows that

Proposition 2.7. *If K is compactly supported, then $K_\varphi^{(n)}$ is square integrable if the function $(K\star\hat K)^n$, where $\hat K(z) = K(-z)$, is integrable.*

We now *define* the nth Wick power $\eta^{\diamond n}$ of η as the random distribution given by only keeping the dominant term in (2.2):

$$\eta^{\diamond n}(\varphi) = I^{(n)}(K_\varphi^{(n)})\;.$$

By Proposition 2.8, this makes sense as soon as $K\star\hat K \in L^n(\mathbf{R}^{d+1})$. One then has the following result, a version of which can be found for example in [14].

Proposition 2.8. *Let K be a compactly supported kernel of order $\beta \in (\frac{d+2}{2}(1-\frac{1}{n}), \frac{d+2}{2})$ and let $\eta = K\star\xi$ as above. Then, $\eta^{\diamond n}$ is well-defined and belongs almost surely to $\mathcal{C}^\alpha$ for every $\alpha < (2\beta-d-2)\frac{n}{2}$.*

Proof. A simple calculation shows that

$$\big|\big(K\star\hat K\big)(z)\big|^n \lesssim |z|^{(2\beta-d-2)n}\;,$$

so that $\|K_{\varphi_\varepsilon^\lambda}^{(n)}\|_{L^2}^2 \lesssim \lambda^{(2\beta-d-2)n}$. The claim then follows from Theorem 2.5, noting that random variables belonging to a Wiener-Itô chaos of finite order satisfy the equivalence of moments. □

It is important to note that this result is stable: replacing K by a smoothened kernel K_ε and letting $\varepsilon\to 0$ yields convergence in probability of $\eta_\varepsilon^{\diamond n}$ to $\eta^{\diamond n}$ in $\mathcal{C}^\alpha$ (with α as in the statement of the proposition) for most "reasonable" choices of K_ε. Furthermore, for fixed $\varepsilon > 0$, one has an explicit formula relating $\eta_\varepsilon^{\diamond n}$ to η_ε:

$$\eta_\varepsilon^{\diamond n}(z) = H_n(\eta_\varepsilon(z), C_\varepsilon)\;, \tag{2.3}$$

where the rescaled Hermite polynomials $H_n(\cdot, C)$ are related to the standard Hermite polynomials by $H_n(u,C) = C^{n/2}H_n(C^{-1/2}u)$ and we have set $C_\varepsilon = \int K_\varepsilon^2(z)\,dz$.

3. General methodology

The general methodology for providing a robust meaning to equations of the type presented in the introduction is as follows. We remark that the main reason why these equations seem to be ill-posed is that there is no canonical way of multiplying arbitrary distributions. The distributions appearing in our setting are however not arbitrary. For instance, one would expect solutions to semilinear equations of this type to locally "look like" the solutions to the corresponding linear problems. This is because, unlike hyperbolic or dispersive equations, parabolic (or elliptic) equations to not transport singularities. This gives hope that if one could somehow make sense of the nonlinearity, when applied to the solution to the linearised equation (which is a Gaussian process and therefore amenable to explicit calculations), then one could maybe give meaning to the equations themselves.

3.1. The Da Prato-Debussche trick. In some situations, one can apply this idea directly, and this was originally exploited in the series of articles [10–12]. Let us focus on the example of the dynamical Φ^4 model in dimension 2, which is formally given by

$$\partial_t \Phi = \Delta \Phi + C\Phi - \Phi^3 + \xi \, ,$$

where ξ is (spatially periodic) space-time white noise in space dimension 2.

Let now ξ_ε denote a smoothened version of ξ given for example by $\xi_\varepsilon = \rho_\varepsilon \star \xi$, where $\rho_\varepsilon(t,x) = \varepsilon^{-4}\rho(\varepsilon^{-2}t, \varepsilon^{-1}x)$, for some smooth compactly supported space-time mollifier ρ. In this case, denoting again by K a cut-off version of the heat kernel and noting that K is of order 2 (and therefore also of every order less than 2), it is immediate that $\eta = K \star \xi$ satisfies the assumptions of Proposition 2.8 for every integer n.

In view of (2.3), this suggests that it might be possible to show that the solutions to

$$\begin{aligned} \partial_t \Phi_\varepsilon &= \Delta \Phi_\varepsilon + 3C_\varepsilon \Phi_\varepsilon - \Phi_\varepsilon^3 + \xi_\varepsilon \\ &= \Delta \Phi_\varepsilon - H_3(\Phi_\varepsilon, C_\varepsilon) + \xi_\varepsilon \, , \end{aligned} \tag{3.1}$$

with $C_\varepsilon = \int K_\varepsilon^2(z)\, dz$ as above, where $K_\varepsilon = \rho_\varepsilon \star K$, converge to a distributional limit as $\varepsilon \to 0$. This is indeed the case, and the argument goes as follows. Writing $\eta_\varepsilon = K_\varepsilon \star \xi$ and $v_\varepsilon = \Phi_\varepsilon - \eta_\varepsilon$ with Φ_ε the solution to (3.1), we deduce that v_ε solves the equation

$$\partial_t v_\varepsilon = \Delta v_\varepsilon - H_3(\eta_\varepsilon + v_\varepsilon, C_\varepsilon) + R_\varepsilon \, ,$$

for some smooth function R_ε that converges to a smooth limit R as $\varepsilon \to 0$. We then use elementary properties of Hermite polynomials to rewrite this as

$$\begin{aligned} \partial_t v_\varepsilon &= \Delta v_\varepsilon - \big(H_3(\eta_\varepsilon, C_\varepsilon) + 3v_\varepsilon H_2(\eta_\varepsilon, C_\varepsilon) + 3v_\varepsilon^2 \, \eta_\varepsilon + v_\varepsilon^3\big) + R_\varepsilon \\ &= \Delta v_\varepsilon - \big(\eta_\varepsilon^{\diamond 3} + 3v_\varepsilon \eta_\varepsilon^{\diamond 2} + 3v_\varepsilon^2 \, \eta_\varepsilon + v_\varepsilon^3\big) + R_\varepsilon \, . \end{aligned}$$

By Proposition 2.8 (and the remarks that follow), we see that $\eta_\varepsilon^{\diamond n}$ converges in probability to a limit $\eta^{\diamond n}$ in every space $\mathcal{C}^\alpha$ for $\alpha < 0$. We can then *define* a random distribution Φ by $\Phi = \eta + v$, where v is the solution to

$$\partial_t v = \Delta v - \big(\eta^{\diamond 3} + 3v\eta^{\diamond 2} + 3v^2 \, \eta + v^3\big) + R \, . \tag{3.2}$$

As a consequence of Theorem 2.3 (combined with additional estimates showing that the $\mathcal{C}^\gamma$-norm of $K \star (f \mathbf{1}_{t>0})$ is small over short times provided that $f \in \mathcal{C}^\alpha$ for $\alpha \in (-2, 0)$

and $\gamma < \alpha + \beta$), it is relatively easy to show that (3.2) has local solutions, and that these solutions are robust with respect to approximations of $\eta^{\diamond n}$ in $\mathcal{C}^\alpha$ for α sufficiently close to 0. In particular, this shows that one has $\Phi_\varepsilon \to \Phi$ in probability, at least locally in time for short times.

Remark 3.1. The dynamical Φ^4 model in dimension 2 was previously constructed in [1] (see also the earlier work [31] where a related but different process was constructed), but that construction relied heavily on *a priori* knowledge about its invariant measure and it was not clear how robust the construction was with respect to perturbations.

3.2. Breakdown of the argument and a strategy to rescue it. While the argument outlined above works very well for a number of equations, it unfortunately breaks down for the equations mentioned in the introduction. Indeed, consider again (1.4a), but this time in space dimension $d = 3$. In this case, one has $\eta \in \mathcal{C}^{-\frac{1}{2}-\kappa}$ for every $\kappa > 0$ and, by Proposition 2.8, one can still make sense of $\eta^{\diamond n}$ for $n < 5$. One could therefore hope to define again a solution Φ by setting $\Phi = \eta + v$ with v the solution to (3.2). Unfortunately, this is doomed to failure: since $\eta^{\diamond 3} \in \mathcal{C}^{-\frac{3}{2}-\kappa}$ (but no better), one can at best hope to have $v \in \mathcal{C}^{\frac{1}{2}-\kappa}$. As a consequence, both products $v \cdot \eta^{\diamond 2}$ and $v^2 \cdot \eta$ fall outside of the scope of Theorem 2.2 and we cannot make sense of (3.2).

One might hope at this stage that the Da Prato-Debussche trick could be iterated to improve things: identify the "worst" term in the right hand side of (3.2), make sense of it "by hand", and try to obtain a well-posed equation for the remainder. While this strategy can indeed be fruitful and allows us to deal with slightly more singular problems, it turns out to fail in this situation. Indeed, no matter how many times we iterate this trick, the right hand side of the equation for the remainder v will *always* contain a term proportional to $v \cdot \eta^{\diamond 2}$. As a consequence, one can *never* hope to obtain a remainder of regularity better than $\mathcal{C}^{1-\kappa}$ which, since $\eta^{\diamond 2} \in \mathcal{C}^{-1-\kappa}$, shows that it is not possible to obtain a well-posed equation by this method. See also Remark 5.17 below for a more systematic explanation of when this trick fails.

In some cases, one does not even know how to get started: consider the class of "classical" one-dimensional stochastic PDEs given by

$$\partial_t u = \partial_x^2 u + f(u) + g(u)\xi \,, \tag{3.3}$$

where ξ denotes space-time white noise, f and g are fixed smooth functions from $\mathbf{R}$ to $\mathbf{R}$, and the spatial variable x takes values on the circle. Then, we know in principle how to use Itô calculus to make sense of (3.3) by rewriting it as an integral equation and interpreting the integral against ξ as an Itô integral, see [13]. However, this notion of solution is not very robust under approximations since space-time regularisations of the driving noise ξ typically destroy the probabilistic structure required for Itô integration. This is in contrast to the solution theory sketched in Section 3.1 which was very stable under approximations of the driving noise, even though it required suitable adjustments to the equation itself. Unfortunately, the argument of Section 3.1 (try to find some function / distribution η so that $v = u - \eta$ has better regularity properties and then obtain a well-posed equation for v) appears to break down completely.

The main idea now is that even though we may not be able to find a global object η so that $u - \eta$ has better regularity, it might be possible to find a *local* object that does the trick at any one point. More precisely, setting $\eta = K \star \xi$ as above (this time η is a Hölder continuous

function in $\mathcal{C}^{\frac{1}{2}-\kappa}$ for every $\kappa > 0$ by Theorems 2.3 and 2.5), one would expect solutions to (3.3) to be well approximated by

$$u(z') \approx u(z) + g(u(z))\big(\eta(z') - \eta(z)\big)\,. \tag{3.4}$$

The intuition is that since K is regular everywhere except at the origin, convolution with K is "almost" a local operator, modulo more regular parts. Since, near any fixed point z, we would expect $g(u)\xi$ to "look like" $g(u(z))\xi$ this suggests that near that point z, the function $K \star (g(u)\xi)$ should "look like" $g(u(z))\eta$, which is what (3.4) formalises.

Note that this looks very much like a first-order Taylor expansion, but with $\eta(z') - \eta(z)$ playing the role of the linear part $z' - z$. If we assume that (3.4) yields a good approximation to u, then one would also expect that

$$g(u(z')) \approx g(u(z)) + g'(u(z))g(u(z))\big(\eta(z') - \eta(z)\big)\,,$$

so that $g(u)$ has again a "first-order Taylor expansion" of the same type as the one for u. One could then hope that if we know somehow how to multiply η with ξ, this knowledge could be leveraged to define the product between $g(u)$ and ξ in a robust way. It turns out that this is *not* quite enough for the situation considered here. However, this general strategy turns out to be very fruitful, provided that we also control higher-order local expansions of u, and this is precisely what the theory of regularity structures formalises [23, 26]. In particular, besides being applicable to (3.3), it also applies to all of the equations mentioned in the introduction.

4. Regularity structures

We now describe a very general framework in which one can formulate "Taylor expansions" of the type (3.4). We would like to formalise the following features of Taylor expansions. First, the coefficients of a Taylor expansion (i.e. the value and derivatives of a given function in the classical case or the coefficients $u(z)$ and $g(u(z))$ in the case (3.4)) correspond to terms of different degree / homogeneity and should therefore naturally be thought of as elements in some graded vector space. Second, an expansion around a given point can be reexpanded around a different point at the expense of changing coefficients, like so:

$$\begin{aligned} a \cdot 1 + b \cdot x + c \cdot x^2 &= \big(a + bh + ch^2\big) \cdot 1 + \big(b + 2ch\big) \cdot (x - h) + c \cdot (x - h)^2\,, \\ u \cdot 1 + g(u) \cdot \big(\eta(z') - \eta(z)\big) &= \big(u + g(u)(\eta(z'') - \eta(z))\big) \cdot 1 + g(u) \cdot \big(\eta(z') - \eta(z'')\big)\,. \end{aligned}$$

Lastly, we see from these expressions that if we order coefficients by increasing homogeneity, then the linear transformation performing the reexpansion has an upper triangular structure with the identity on the diagonal.

4.1. Basic definitions. The properties just discussed are reflected in the following algebraic structure.

Definition 4.1. A regularity structure $\mathscr{T} = (A, T, G)$ consists of the following elements:

1. A discrete index set $A \subset \mathbf{R}$ such that $0 \in A$ and A is bounded from below.
2. A *model space* $T = \bigoplus_{\alpha \in A} T_\alpha$, with each T_α a Banach space; elements in T_α are said to have *homogeneity* α. Furthermore T_0 is one-dimensional and has a distinguished basis vector $\mathbf{1}$. Given $\tau \in T$, we write $\|\tau\|_\alpha$ for the norm of its component in T_α.

3. A *structure group* G of (continuous) linear operators acting on T such that, for every $\Gamma \in G$, every $\alpha \in A$, and every $\tau_\alpha \in T_\alpha$, one has

$$\Gamma\tau_\alpha - \tau_\alpha \in T_{<\alpha} := \bigoplus_{\beta<\alpha} T_\beta \, . \tag{4.1}$$

Furthermore, $\Gamma\mathbf{1} = \mathbf{1}$ for every $\Gamma \in G$.

The prime example of a regularity structure one should keep in mind is the one associated to Taylor polynomials on space-time $\mathbf{R}^{d+1}$. In this case, the space T is given by all polynomials in $d+1$ indeterminates $X_0, \dots, X_d$, with X_0 representing the "time" coordinate. It comes with a canonical basis given by all monomials of the type $X^k = X_0^{k_0} \cdots X_d^{k_d}$ with k an arbitrary multiindex. The basis vector $\mathbf{1}$ is the one corresponding to the zero multiindex. The space T has a natural grading by postulating that the homogeneity of X^k is $|k| = 2k_0 + \sum_{i\neq 0} k_i$ and a natural norm by postulating that $\|X^k\| = 1$. In the case of the polynomial regularity structure, the structure group G is simply given by $\mathbf{R}^{d+1}$, endowed with addition, and acting on monomials by

$$\hat\Gamma_h X^k = (X-h)^k = (X_0 - h_0)^{k_0} \cdots (X_d - h_d)^{k_d} \, . \tag{4.2}$$

It is immediate that all axioms of a regularity structure are satisfied in this case.

In the case of polynomials, there is a natural "realisation" of the structure $\mathscr{T}$ at each space-time point z, which is obtained by turning an abstract polynomial into the corresponding concrete polynomial (viewed now as a real-valued function on $\mathbf{R}^{d+1}$) based at z. In other words, we naturally have a family of linear maps $\Pi_z \colon T \to \mathcal{C}^\infty(\mathbf{R}^d)$ given by

$$\big(\Pi_z X^k\big)(z') = (z'_0 - z_0)^{k_0} \cdots (z'_d - z_d)^{k_d} \, . \tag{4.3}$$

It is immediate that the group G transforms these maps into each other in the sense that $\Pi_z \hat\Gamma_h = \Pi_{z+h}$. It is furthermore an immediate consequence of the scaling properties of monomials that the maps Π_z and the representation $h \mapsto \hat\Gamma_h$ of $\mathbf{R}^{d+1}$ are "compatible" with our grading for the model space T. More precisely, one has

$$\langle \varphi_z^\lambda, \Pi_z X^k \rangle = \lambda^{|k|} \langle \varphi, \Pi_0 X^k \rangle \, , \qquad \|\hat\Gamma_h X^k\|_\ell = C_{k,\ell} |h|^{|k|-\ell} \, ,$$

for some constants $C_{k,\ell}$ and every $\ell \le |k|$. Here, $\langle \cdot, \cdot \rangle$ denotes again the usual L^2-scalar product.

These observations suggest the following definition of a "model" for $\mathscr{T}$, where we impose properties similar to the ones we just found for the polynomial model. A model always requires the specification of an ambient space, together with a possibly inhomogeneous scaling. For definiteness, we will fix our ambient space to be $\mathbf{R}^{d+1}$ endowed with the parabolic scaling as above. We also denote by $\mathcal{S}'$ the space of all distributions (the letter $\mathcal{D}$ is reserved for a different usage below). We also denote by $L(E,F)$ the set of all continuous linear maps between the topological vector spaces E and F.

Definition 4.2. Given a regularity structure $\mathscr{T}$, a *model* for $\mathscr{T}$ consists of maps

$$\mathbf{R}^{d+1} \ni z \mapsto \Pi_z \in L(T, \mathcal{S}') \, , \qquad \mathbf{R}^{d+1} \times \mathbf{R}^{d+1} \ni (z, z') \mapsto \Gamma_{zz'} \in G \, ,$$

satisfying the algebraic compatibility conditions

$$\Pi_z \Gamma_{zz'} = \Pi_{z'} \, , \qquad \Gamma_{zz'} \circ \Gamma_{z'z''} = \Gamma_{zz''} \, , \tag{4.4}$$

as well as the analytical bounds

$$|\langle \Pi_z \tau, \varphi_z^\lambda \rangle| \lesssim \lambda^\alpha \|\tau\| \,, \qquad \|\Gamma_{zz'}\tau\|_\beta \lesssim |z-z'|^{\alpha-\beta}\|\tau\| \,. \tag{4.5}$$

Here, the bounds are imposed uniformly over all $\tau \in T_\alpha$, all $\beta < \alpha \in A$, and all test functions $\varphi \in \mathcal{B}_r$ with $r = \inf A$, and locally uniformly in z and z'.

Remark 4.3. These definitions suggest a natural topology for the space $\mathscr{M}$ of all models for a given regularity structure, generated by the following family of pseudo-metrics indexed by compact sets K:

$$\sup_{z\in K}\Big(\sup_{\varphi,\lambda,\alpha,\tau} \lambda^{-\alpha}|\langle \Pi_z\tau - \bar\Pi_z\tau, \varphi_z^\lambda\rangle| + \sup_{|z-z'|\le 1}\sup_{\alpha,\beta,\tau} |z-z'|^{\beta-\alpha}\|\Gamma_{zz'}\tau - \bar\Gamma_{zz'}\tau\|_\beta\Big) \,. \tag{4.6}$$

Here the inner suprema run over the same sets as before, but with $\|\tau\| = 1$.

4.2. Hölder classes. It is clear from the above discussion that if $\mathscr{T}$ is the polynomial structure, Π is defined as in (4.3), and $\Gamma_{zz'} = \hat\Gamma_{z'-z}$ with $\hat\Gamma_h$ as in (4.2), then (Π,Γ) is a model for $\mathscr{T}$ in the sense of Definition 4.2. Given an arbitrary regularity structure $\mathscr{T}$ and an arbitrary model (Π,Γ), it is now natural to define the corresponding "Hölder spaces" as spaces of distributions that can locally (near any space-time point z) be approximated by $\Pi_z\tau$ for some $\tau \in T$. This would be the analogue to the statement that a smooth function is one that can locally be approximated by a polynomial.

There is however one major difference with the case of smooth functions. It is of course the case that if f is smooth, then the coefficients of the Taylor expansion of f at any point are uniquely determined by the behaviour of f in the vicinity of that point. This is in general *not* the case anymore in the context of the framework we just described. To appreciate this fact, consider the following example. Fix $\alpha \in (0,1)$ and $m \in \mathbf{N}$, and take for $\mathscr{T}$ the regularity structure where $A = \{0,\alpha\}$, $T_0 \cong \mathbf{R}$ with basis vector $\mathbf{1}$, $T_\alpha \cong \mathbf{R}^m$ with basis vectors $(e_i)_{i\le m}$, and structure group $G \cong \mathbf{R}^m$ acting on T via $\hat\Gamma_h e_i = e_i - h_i\mathbf{1}$. Let then W be an $\mathbf{R}^m$-valued α-Hölder continuous function defined on the ambient space and set

$$\Pi_z\mathbf{1} = 1 \,, \qquad \big(\Pi_z e_i\big)(z') = W_i(z') - W_i(z) \,, \qquad \Gamma_{zz'} = \hat\Gamma_{W(z)-W(z')} \,.$$

Again, it is straightforward to verify that this does indeed define a model for $\mathscr{T}$. In fact, setting $m = 1$ and $W = \eta$, this is precisely the structure one would use to formalise the expansion (3.4).

Let now $F\colon \mathbf{R}^m \to \mathbf{R}$ be a smooth function and consider the function f on the ambient space given by $f(z) = F(W(z))$. For any z, we furthermore set

$$T \ni \hat f(z) = F(W(z))\,\mathbf{1} + \sum_{i=1}^m (\partial_i F)(W(z))\, e_i \,.$$

It then follows immediately from the usual Taylor expansion of F and the definition of the model (Π,Γ) that one has the bound

$$\big|f(z') - \big(\Pi_z \hat f(z)\big)(z')\big| \lesssim |z-z'|^{2\alpha} \,, \tag{4.7}$$

so that in this context and with respect to this specific model, the function f behaves as if it were of class $\mathcal{C}^{2\alpha}$ with "Taylor series" given by $\hat f$. In the case where the underlying

space is one-dimensional, this is precisely the insight exploited in the theory of rough paths [16, 35, 36] in order to develop a pathwise approach to stochastic calculus. More specifically, the perspective given here (i.e. controlling functions via analogues to Taylor expansion) is that of the theory of controlled rough paths developed in [18].

It is now very natural to ask whether, just like in the case of smooth functions, a bound of the type (4.7) is sufficient to uniquely specify $\hat{f}(z)$ for every point z. Unfortunately, the answer to this question is that "it depends". The reason is that while (4.5) imposes an upper bound on the behaviour of Π_z in the vicinity of z, it does *not* impose any corresponding lower bound. For example, $W \equiv 0$ is an α-Hölder continuous function that we could have used to build our model. In that case, the value of the e_i-component in $\hat{f}$ is completely irrelevant for (4.7), so that uniqueness of the "Taylor series" fails. Suppose on the other hand that the underlying space is one-dimensional, that $\alpha \in (\frac{1}{4}, \frac{1}{2})$, and that W is a typical sample path of a Brownian trajectory. In this case it was shown in [27, Thm 3.4] that a bound of the type (4.7) is indeed sufficient to uniquely determine all the coefficients of $\hat{f}$ (at least for almost all Brownian trajectories).

Remark 4.4. The fact that $\hat{f}$ is uniquely determined by f in the Brownian case can be interpreted as an analogue to the fact that the Doob-Meyer decomposition of a semimartingale is unique. Since the statement given in [27] is quantitative, it can be interpreted as a deterministic analogue to Norris's lemma, of which various incarnations can be found in [6, 33, 38].

Consider now a sequence W^ε of smooth (random) functions so that W^ε converges to Brownian motion in $\mathcal{C}^\alpha$ as $\varepsilon \to 0$. For definiteness, take for W_ε piecewise linear interpolations on a grid of size ε. Then, if we know *a priori* that we have a bound of the type (4.7) with a proportionality constant of order 1, this determines the coefficients of $\hat{f}$ "almost uniquely" up to an error of order about $\varepsilon^{2\alpha-\frac{1}{2}}$.

What this discussion suggests is that we should really reverse our point of view from what we are used to: instead of fixing a function and asking whether it has a certain Hölder regularity by checking whether it is possible to find a "Taylor expansion" at each point satisfying a bound of the type (4.7), we should take the candidate expansion as our fundamental object and ask under which condition it does indeed approximate one single function / distribution around each point at the prescribed order. More precisely, fix some $\gamma > 0$ (the order of our "Taylor expansion") and consider a function $f\colon \mathbf{R}^{d+1} \to T_{<\gamma}$. Under which assumptions can we find a distribution ζ such that ζ "looks like" the distribution $\Pi_z f(z)$ (in a suitable sense) near every point z? We claim that the "right" answer is given by the following definition.

Definition 4.5. Given a regularity structure $\mathscr{T}$ and a model (Π, Γ) as above, we define $\mathcal{D}^\gamma$ as the space of functions $f\colon \mathbf{R}^{d+1} \to T_{<\gamma}$ such that the bound

$$\|f(z) - \Gamma_{zz'} f(z')\|_\alpha \lesssim |z - z'|^{\gamma-\alpha} \,. \tag{4.8}$$

holds for every $\alpha < \gamma$, locally uniformly in z and z'.

Remark 4.6. This definition makes sense and is non-empty even for negative γ, as long as $\gamma > \inf A$.

Remark 4.7. The notation $\mathcal{D}^\gamma$ is really an abuse of notation, since even for a given regularity structure there isn't one single space $\mathcal{D}^\gamma$, but a whole collection of them, one for each model $(\Pi, \Gamma) \in \mathscr{M}$. More formally, one should really consider the space $\mathscr{M} \ltimes \mathcal{D}^\gamma$ consisting of

pairs $((\Pi,\Gamma),f)$ such that f belongs to the space $\mathcal{D}^\gamma$ based on the model (Π,Γ). The space $\mathscr{M}\ltimes\mathcal{D}^\gamma$ also comes with a natural topology.

In the case where $\mathscr{T}$ is the polynomial regularity structure and (Π,Γ) are the usual Taylor polynomials as above, one can see that this definition coincides with the usual definition of $\mathcal{C}^\gamma$ (except at integer values where $\mathcal{D}^1$ describes Lipschitz continuous functions, etc). In this case, the component $f_0(z)=\langle\mathbf{1},f(z)\rangle$ of $f(z)$ in T_0 (here we write $\langle\mathbf{1},\cdot\rangle$ for the basis element of T^* dual to $\mathbf{1}$) is the only reasonable candidate for the function represented by f. Furthermore, $\langle\mathbf{1},\Gamma_{zz'}f(z')\rangle$ is nothing but the candidate Taylor expansion of f around z', evaluated at z. The bound (4.8) with $\alpha=0$ is then just a statement of the fact that f_0 is of class $\mathcal{C}^\gamma$ and that $f(z)$ is its Taylor series of order γ at z. The corresponding bounds for $\alpha>0$ then follow immediately, since they merely state that the αth derivative of f_0 is of class $\mathcal{C}^{\gamma-\alpha}$.

4.3. The reconstruction operator. The situation is much less straightforward when the model space T contains components of negative homogeneity. In this case, the bounds (4.5) allow the model Π_z to consist of genuine distributions and we do not anymore have an obvious candidate for the distribution represented by f. The following result shows that such a distribution nevertheless always exists and is unique as soon as $\gamma>0$. This also provides an *a posteriori* justification for our definition of the spaces $\mathcal{D}^\gamma$.

Theorem 4.8. *Consider a regularity structure $\mathscr{T}=(A,T,G)$ and fix $\gamma>r=\inf A$. Then, there exists a continuous map $\mathcal{R}\colon\mathscr{M}\ltimes\mathcal{D}^\gamma\to\mathcal{S}'$ (the "reconstruction map") with the property that*

$$\big|\big(\mathcal{R}(\Pi,\Gamma,f)-\Pi_z f(z)\big)(\varphi_z^\lambda)\big|\lesssim\lambda^\gamma\,, \tag{4.9}$$

uniformly over $\lambda\in(0,1]$ and $\varphi\in\mathcal{B}_r$, and locally uniformly over $z\in\mathbf{R}^{d+1}$. Furthermore, for any given model (Π,Γ), the map $f\mapsto\mathcal{R}(\Pi,\Gamma,f)$ is linear. If $\gamma>0$, the map $\mathcal{R}$ is uniquely specified by the requirement (4.9).

Remark 4.9. In the sequel, we will always consider (Π,Γ) as fixed and view $\mathcal{R}$ as a linear map, writing $\mathcal{R}f$ instead of $\mathcal{R}(\Pi,\Gamma,f)$. The above notation does however make it plain that the full map $\mathcal{R}$ is not a linear map.

Remark 4.10. An important special case is given by situations where $\Pi_z\tau$ happens to be a continuous function for every $\tau\in T$ and every z. Then, it turns out that $\mathcal{R}f$ is also a continuous function and one simply has

$$\big(\mathcal{R}f\big)(z)=\big(\Pi_z f(z)\big)(z)\,. \tag{4.10}$$

In the general case, this formula makes of course no sense since $\Pi_z f(z)$ is a distribution and cannot be evaluated at z.

Remark 4.11. We made a slight abuse of notation here since there is really a family of operators $\mathcal{R}^\gamma$, one for each regularity. However, this abuse is justified by the following consistency relation. Given $f\in\mathcal{D}^\gamma$ and $\tilde\gamma<\gamma$, one can always construct $\tilde f$ by projecting $f(z)$ onto $T_{<\tilde\gamma}$ for every z. It turns out that one then necessarily has $\tilde f\in\mathcal{D}^{\tilde\gamma}$ and $\mathcal{R}\tilde f=\mathcal{R}f$, provided that $\tilde\gamma>0$. This is also consistent with (4.10) since, if $\Pi_z\tau$ is a continuous function and the homogeneity of τ is strictly positive, then $\big(\Pi_z\tau\big)(z)=0$.

We refer to [23, Thm 3.10] for a full proof of Theorem 4.8 and to [22] for a simplified proof that only gives continuity in each "fiber" $\mathcal{D}^\gamma$. The main idea is to use a basis of compactly supported wavelets to construct approximations $\mathcal{R}^n$ in such a way that our definitions can be exploited in a natural way to compare $\mathcal{R}^{n+1}$ with $\mathcal{R}^n$ and show that the sequence of approximations is Cauchy in a suitable space of distributions $\mathcal{C}^\alpha$. In the most important case when $\gamma > 0$, it turns out that while the existence of a map $\mathcal{R}$ with the required properties is highly non-trivial, its uniqueness is actually quite easy to see. If $\gamma \leq 0$ on the other hand, it is clear that $\mathcal{R}$ cannot be uniquely determined by (4.9), since this bound remains unchanged if we add to $\mathcal{R}$ any distribution in $\mathcal{C}^\gamma$. The existence of $\mathcal{R}$ in the case $\gamma < 0$ is however still a non-trivial result since in general one has $\mathcal{R}f \notin \mathcal{C}^\gamma$!

5. Regularity structures for SPDEs

We now return to the problem of providing a robust well-posedness theory for stochastic PDEs of the type (1.2), (1.4), (1.3), or even just (3.3). Our aim is to build a suitable regularity structure for which we can reformulate our SPDE as a fixed point problem in $\mathcal{D}^\gamma$ for a suitable value of γ.

Remark 5.1. Actually, it turns out that since we are interested in Cauchy problems, there will always be some singularity at $t = 0$. This introduces additional technical complications which we do not wish to dwell upon.

5.1. General construction of the model space. Our first task is to construct the model space T. Since we certainly want to be able to represent arbitrary smooth functions (for example in order to be able to take into account the contribution of the initial condition), we want T to contain the space $\bar{T}$ of abstract polynomials in $d + 1$ indeterminates endowed with the parabolic grading described in Section 4.1. Since the noise ξ cannot be adequately represented by polynomials, we furthermore add a basis vector Ξ to T, which we postulate to have some homogeneity $\alpha < 0$ such that $\xi \in \mathcal{C}^\alpha$. In the case of space-time white noise, we would choose $\alpha = -\frac{d}{2} - 1 - \kappa$ for some (typically very small) exponent $\kappa > 0$.

At this stage, the discussion following (3.4) suggests that if our structure T contains a basis vector τ of homogeneity β representing some distribution η involved in the description of the right hand side of our equation, then it should also contain a basis vector of homogeneity $\beta + 2$ (the "2" here comes from the fact that convolution with the heat kernel yields a gain of 2 in regularity) representing the distribution $K \star \eta$ involved in the description of the solution to the equation. Let us denote this new basis vector by $\mathcal{I}(\tau)$, where $\mathcal{I}$ stands for "integration". In the special case where $\tau \in \bar{T}$, so that it represents an actual polynomial, we do not need any new symbol since K convolved with a polynomial yields a smooth function. One way of formalising this is to simply postulate that $\mathcal{I}(X^k) = 0$ for every multiindex k.

Remark 5.2. For consistency, we will also always assume that $\int K(z)Q(z)\,dz = 0$ for all polynomials Q of some fixed, but sufficiently high, degree. Since K is an essentially arbitrary truncation of the heat kernel, we can do this without loss of generality.

If the right hand side of our equation involves the spatial derivatives of the solution, then, for each basis vector τ of homogeneity β representing some distribution η appearing in the description of the solution, we should also have a basis vector $\mathscr{D}_i\tau$ of homogeneity

$\beta - 1$ representing $\partial_i \eta$ and appearing in the description of the derivative of the solution in the direction x_i.

Finally, if the right hand side of our equation involves a product between two terms F and $\bar{F}$, and if basis vectors τ and $\bar{\tau}$ respectively are involved in their description, then we should also have a basis vector $\tau\bar{\tau}$ which would be involved in the description of the product. If τ and $\bar{\tau}$ represent the distributions η and $\bar{\eta}$ respectively, then this new basis vector represents the distribution $\eta\bar{\eta}$, whatever this actually means. Regarding its homogeneity, by analogy with the case of polynomials, it is natural to impose that the homogeneity of $\tau\bar{\tau}$ is the sum of the homogeneities of its two factors.

This suggests that we should build T by taking as its basis vectors some formal expressions built from the symbols X and Ξ, together with the operations $\mathcal{I}(\cdot)$, $\mathscr{D}_i$, and multiplication. Furthermore, the natural way of computing the homogeneity of a formal expression in view of the above is to associate homogeneity 2 to X_0, 1 to X_i for $i \neq 0$, α to Ξ, 2 to $\mathcal{I}(\cdot)$, and -1 to $\mathscr{D}_i$, and to simply add the homogeneities of all symbols appearing in any given expression. Denote by $\mathcal{F}$ the collection of all formal expressions that can be constructed in this way and denote by $|\tau|$ the homogeneity of $\tau \in \mathcal{F}$, so we have for example

$$\bigl|X_i\Xi\bigr| = \alpha + 1\,, \qquad \bigl|\mathcal{I}(\Xi)^2\mathcal{I}(X_i\mathscr{D}_j\mathcal{I}(\Xi))\bigr| = 3\alpha + 8\,, \qquad \text{etc.}$$

We note however that if we simply took for T the space of linear combinations of *all* elements in $\mathcal{F}$ then, since $\alpha < 0$, there would be basis vectors of arbitrarily negative homogeneity, which would go against Definition 4.1. What saves us is that most formal expressions are not needed in order to formulate our equations as fixed point problems. For example, the expression Ξ^2 is useless since we would never try to square the driving noise. Similarly, if we consider (1.4a), then $\mathcal{I}(\Xi)$ is needed for the description of the solution, which implies that $\mathcal{I}(\Xi)^2$ and $\mathcal{I}(\Xi)^3$ are needed to describe the right hand side, but we do not need $\mathcal{I}(\Xi)^4$ for example.

5.2. Specific model spaces. This suggests that we should take T as the linear combinations of only those formal expressions $\tau \in \mathcal{F}$ that are actually expected to appear in the description of the solution to our equation or its right hand side. Instead of trying to formulate a general construction (see [23, Sec. 8.1] for such an attempt), let us illustrate this by a few examples. We first focus on the case of the KPZ equation (1.3) and we construct subsets $\mathcal{U}$ and $\mathcal{V}$ of $\mathcal{F}$ that are used in the description of the solution and the right hand side of the equation respectively. These are defined as the smallest subsets of $\mathcal{F}$ with the following properties:

$$\mathcal{T} \subset \mathcal{U} \cap \mathcal{V}\,, \quad \{\mathcal{I}(\tau) : \tau \in \mathcal{V} \setminus \mathcal{T}\} \subset \mathcal{U}\,, \quad \{\Xi\} \cup \{\mathscr{D}\tau_1 \cdot \mathscr{D}\tau_2 : \tau_i \in \mathcal{U}\} \subset \mathcal{V}\,. \tag{5.1}$$

where we used the notation $\mathcal{T} = \{X^k\}$ with k running over all multiindices, so that the space of Taylor polynomials $\bar{T}$ is the linear span of $\mathcal{T}$. We then define T as the space of all linear combinations of elements of $\mathcal{U} \cup \mathcal{V}$. We also denote by $T_{\mathcal{U}}$ the subspace of T spanned by $\mathcal{U}$. This construction is such that if we have any function $H\colon \mathbf{R}^{d+1} \to T_{\mathcal{U}}$, then we can define in a natural way a function $\Xi - (\mathscr{D}H)^2\colon \mathbf{R}^{d+1} \to T$ by the last property. Furthermore, by the second property, one has again $\mathcal{I}(\Xi - (\mathscr{D}H)^2)\colon \mathbf{R}^{d+1} \to T_{\mathcal{U}}$, which suggests that T is indeed sufficiently rich to formulate a fixed point problem mimicking the mild formulation of (1.3). Furthermore, one has

Lemma 5.3. *If $\mathcal{U}$ and $\mathcal{V}$ are the smallest subsets of $\mathcal{F}$ satisfying* (5.1) *and one has $|\Xi| > -2$ then, for every $\gamma > 0$, the set $\{\tau \in \mathcal{U} \cup \mathcal{V} : |\tau| < \gamma\}$ is finite.*

The condition $\alpha > -2$ corresponds to the restriction $d < 2$, which makes sense since 2 is the critical dimension for the KPZ equation [32]. The other example we would like to consider is the class of SPDEs (3.3). In this case, the right hand side is not polynomial. However, we can apply the same methodology as above as if the nonlinear functions f and g were simply polynomials of arbitrary degree. We thus impose $\mathcal{T} \subset \mathcal{U} \cap \mathcal{V}$ and $\{\mathcal{I}(\tau) : \tau \in \mathcal{V} \setminus \mathcal{T}\}$ as before, and then further impose that

$$\Big\{\Xi \prod_{i=1}^m \tau_i : m \geq 1 \,\&\, \tau_i \in \mathcal{U}\Big\} \cup \Big\{\prod_{i=1}^m \tau_i : m \geq 1 \,\&\, \tau_i \in \mathcal{U}\Big\} \subset \mathcal{V}\,.$$

Again, we have $\mathcal{U} \subset \mathcal{V}$ and we define T as before. Furthermore, it is straightforward to verify that the analogue to Lemma 5.3 holds, provided that $|\Xi| > -2$.

5.3. Construction of the structure group. Now that we have some idea on how to construct T for the problems that are of interest to us (with a slightly different construction for each class of models but a clear common thread), we would like to build a corresponding structure group G. In order to give a motivation for the definition of G, it is very instructive to simultaneously think about the structure of the corresponding models. Let us first consider some smooth driving noise, which we call ξ_ε to distinguish it from the limiting noise ξ. At this stage however, this should be thought of as simply a fixed smooth function. In view of the discussion of Section 5.1, for each of the model spaces built in Section 5.2, we can associate to ξ_ε a linear map $\mathbf{\Pi}\colon T \to \mathcal{C}^\infty(\mathbf{R}^{d+1})$ in the following way. We set

$$\big(\mathbf{\Pi} X_i\big)(z) = z_i\,, \qquad \big(\mathbf{\Pi}\Xi\big)(z) = \xi_\varepsilon(z)\,, \tag{5.2a}$$

and we then define $\mathbf{\Pi}$ recursively by

$$\mathbf{\Pi}\mathcal{I}(\tau) = K \star \mathbf{\Pi}\tau\,, \qquad \mathbf{\Pi}\mathscr{D}_i\tau = \partial_i \mathbf{\Pi}\tau\,, \qquad \mathbf{\Pi}(\tau\bar\tau) = \big(\mathbf{\Pi}\tau\big)\cdot\big(\mathbf{\Pi}\bar\tau\big)\,, \tag{5.2b}$$

where $\cdot$ simply denotes the pointwise product between smooth functions. At this stage, it is however not clear how one would build an actual model in the sense of Definition 4.2 associated to ξ_ε. It is natural that one would set

$$\big(\Pi_z X_i\big)(z') = z_i' - z_i\,, \qquad \big(\Pi_z \Xi\big)(z') = \xi_\varepsilon(z')\,, \tag{5.3a}$$

and then

$$\Pi_z \mathscr{D}_i\tau = \partial_i \Pi_z\tau\,, \qquad \Pi_z(\tau\bar\tau) = \big(\Pi_z\tau\big)\cdot\big(\Pi_z\bar\tau\big)\,. \tag{5.3b}$$

It is less clear *a priori* how to define $\Pi_z\mathcal{I}(\tau)$. The problem is that if we simply set $\Pi_z\mathcal{I}(\tau) = K\star\Pi_z\tau$, then the bound (4.5) would typically no longer be compatible with the requirement that $|\mathcal{I}(\tau)| = |\tau| + 2$. One way to circumvent this problem is to simply subtract the Taylor expansion of $K\star\Pi_z\tau$ around z up to the required order. We therefore set

$$\big(\Pi_z\mathcal{I}(\tau)\big)(z') = \big(K\star\Pi_z\tau\big)(z') - \sum_{|k|<|\tau|+2} \frac{(z'-z)^k}{k!}\big(D^{(k)}K\star\Pi_z\tau\big)(z)\,. \tag{5.3c}$$

It can easily be verified (simply proceed recursively) that if we define Π_z in this way and $\mathbf{\Pi}$ as in (5.2) then, for every z, one can find a linear map $F_z\colon T \to T$ such that $\Pi_z = \mathbf{\Pi}F_z$. In particular, one has $\Pi_{z'} = \Pi_z F_z^{-1} F_{z'}$. Furthermore, F_z is "upper triangular"

with the identity on the diagonal in the sense of (4.1). It is also easily seen by induction that the matrix elements of F_z are all given by some polynomials in z and in the quantities $\big(D^{(k)}K \star \Pi_z\tau\big)(z)$.

This suggests that we should take for G the set of all linear maps that can appear in this fashion. It is however not clear in principle how to describe G more explicitly and it is also not clear that it even forms a group. In order to describe G, it is natural to introduce a space T_+ which is given by all possible polynomials in $d+1$ commuting variables $\{Z_i\}_{i=0}^d$ as well as countably many additional commuting variables $\{\mathcal{J}_k(\tau) : \tau \in (\mathcal{U} \cup \mathcal{V}) \setminus \mathcal{T} \;\&\; |k| < |\tau| + 2\}$. One should think of Z_i as representing z_i and $\mathcal{J}_k(\tau)$ as representing $\big(D^{(k)}K \star \Pi_z\tau\big)(z)$, so that the matrix elements of F_z are represented by elements of T_+. There are no relations between these coefficients, which suggests that elements of G are described by an arbitrary morphism $f\colon T_+ \to \mathbf{R}$, i.e. an arbitrary linear map which furthermore satisfies $f(\sigma\bar\sigma) = f(\sigma)\,f(\bar\sigma)$, so that it is uniquely determined by $f(Z_i)$ and $f(\mathcal{J}_k(\tau))$.

Given any linear map $\Delta\colon T \to T \otimes T_+$ and a morphism f as above, one can then define a linear map $\hat\Gamma_f\colon T \to T$ by

$$\hat\Gamma_f\tau = \big(I \otimes f\big)\Delta\tau\;.$$

(Here we identify T with $T \otimes \mathbf{R}$ in the obvious way.) The discussion given above then suggests that it is possible to construct Δ in such a way that if we define f_z by

$$f_z(Z_i) = z_i\;, \qquad f_z(\mathcal{J}_k(\tau)) = \big(D^{(k)}K \star \Pi_z\tau\big)(z)\;, \tag{5.4}$$

then one has $\hat\Gamma_{f_z} = F_z$. The precise definition of Δ is irrelevant for our discussion, but a recursive description of it can easily be recovered simply by comparing (5.3) to (5.2). In particular, it is possible to show that $\Delta\tau$ is of the form

$$\Delta\tau = \tau \otimes \mathbf{1} + \sum_i c_i^\tau \tau_i \otimes \sigma_i\;, \tag{5.5}$$

for some expressions $\tau_i \in T$ with $|\tau_i| < |\tau|$ and for some non-empty monomials $\sigma_i \in T_+$ such that $|\sigma_i| + |\tau|_i = |\tau|$. Here, we associate a homogeneity to elements in T_+ by setting $|Z_0| = 2$, $|Z_i| = 1$ for $i \neq 0$, and $|\mathcal{J}_k(\tau)| = |\tau| + 2 - |k|$.

In particular, we see that if we let $e\colon T_+ \to \mathbf{R}$ be the trivial morphism for which $e(Z_i) = e(\mathcal{J}_k(\tau)) = 0$, so that one only has $e(\mathbf{1}) = 1$ where $\mathbf{1}$ is the empty product, then $\hat\Gamma_e\tau = \tau$. The important fact for our purpose is the following, a proof of which can be found in [23, Sec. 8]. Here, we denote by $\mathcal{M}\colon T_+ \otimes T_+ \to T_+$ the multiplication operator $\mathcal{M}(\sigma\otimes\bar\sigma) = \sigma\bar\sigma$ and by I the identity.

Theorem 5.4. *There exists a map $\Delta^+\colon T_+ \to T_+ \otimes T_+$ such that the following identities hold:*

$$\begin{aligned} \Delta^+(\sigma\bar\sigma) &= \big(\Delta^+\sigma\big)\cdot\big(\Delta^+\bar\sigma\big)\;, \qquad (\Delta\otimes I)\Delta = (I\otimes\Delta^+)\Delta\;,\\ (e\otimes I)\Delta^+ &= (I\otimes e)\Delta^+ = I\;, \quad (\Delta^+\otimes I)\Delta^+ = (I\otimes\Delta^+)\Delta^+\;. \end{aligned} \tag{5.6}$$

Furthermore, there exists a map $\mathcal{A}\colon T_+ \to T_+$ which is multiplicative in the sense that $\mathcal{A}(\sigma\bar\sigma) = (\mathcal{A}\sigma)\cdot(\mathcal{A}\bar\sigma)$, and which is such that $\mathcal{M}(I\otimes\mathcal{A})\Delta^+ = \mathcal{M}(\mathcal{A}\otimes I)\Delta^+ = e$, with $e\colon T_+ \to \mathbf{R}$ as above.

Remark 5.5. In technical lingo, this lemma states that $(T_+, \cdot, \Delta^+)$ is a Hopf algebra with antipode $\mathcal{A}$, and that T is a comodule over T_+.

The importance of this result is that it shows that G is indeed a group. For any two morphisms f and g, we can define a linear map $f\circ g\colon T_+ \to \mathbf{R}$ by $\big(f\circ g\big)(\sigma) = \big(f\otimes g\big)\Delta^+\sigma$. As a consequence of the first identity in (5.6), $f\circ g$ is again a morphism on T_+. As a consequence of the second identity, one has $\hat\Gamma_{f\circ g} = \Gamma_f\,\Gamma_g$. The last identity shows that $(f_1\circ f_2)\circ f_3 = f_1\circ(f_2\circ f_3)$, while the properties of $\mathcal{A}$ ensure that if we set $f^{-1}(\sigma) = f(\mathcal{A}\sigma)$, then $f\circ f^{-1} = f^{-1}\circ f = e$. Finally, the third identity in (5.6) shows that e is indeed the identity element, thus turning the set of all morphisms of T_+ into a group under $\circ$, acting on T via $\hat\Gamma$.

Let us now turn back to our models. Given a smooth function ξ_ε, we define Π_z as in (5.3) and f_z by (5.4). We then also define linear maps $\Gamma_{zz'}$ by $\Gamma_{zz'} = \hat\Gamma_{\gamma_{zz'}}$ with $\gamma_{zz'} = f_z^{-1}\circ f_{z'}$. We then have

Lemma 5.6. *For every smooth function ξ_ε, the pair (Π,Γ) defined above is a model.*

Proof. The algebraic constraints (4.4) are satisfied essentially by definition. The first bound of (4.5) can easily be verified recursively by (5.3). The only non-trivial fact is that the matrix elements of $\Gamma_{zz'}$ satisfy the right bound. If one can show that $|\gamma_{zz'}(\sigma)| \lesssim |z-z'|^{|\sigma|}$, this in turn follows from (5.5). This bound is non-trivial and was obtained in [23, Prop. 8.27]. □

5.4. Admissible models. Thanks to Lemma 5.6, we now have a large class of models for the regularity structures built in the previous two subsections. However, we do not want to restrict ourselves to this class (or even its closure). The reason is that if we define products in the "naïve" way given by the second identity in (5.3b), then there will typically be some situations where the result diverges as we let $\varepsilon\to 0$ in ξ_ε. Therefore, we do not impose this relation in general but rather view it as the *definition* of the product, i.e. we interpret it as

$$\big(\Pi_z\tau\big)\cdot\big(\Pi_z\bar\tau\big) := \Pi_z(\tau\bar\tau)\;.$$

However, the remainder of the structure described in (5.3) is required for X_i, $\mathscr{D}_i$ and $\mathcal{I}$ to have the correct interpretation. This motivates the following definition.

Definition 5.7. Given a regularity structure $\mathscr{T}$ constructed as in Sections 5.2 and 5.3, we say that a model (Π,Γ) is *admissible* if it satisfies $\big(\Pi_z X_i\big)(z') = z'_i - z_i$, $\Pi_z\mathscr{D}_i\tau = \partial_i\Pi_z\tau$, as well as (5.3c) and if furthermore $\Gamma_{zz'} = \hat\Gamma_{f_z}^{-1}\hat\Gamma_{f_{z'}}$ with f_z given by (5.4). We will denote the space of all admissible models by $\mathscr{M}_0\subset\mathscr{M}$.

Remark 5.8. In the particular case of admissible models for a regularity structure of the type considered here, the data of the single linear map $\mathbf{\Pi}$ as above is sufficient to reconstruct the full model (Π,Γ).

Note that at this stage, it is not clear whether this concept is even well-defined: in general, $D^{(k)}K\star\Pi_z\tau$ will be a distribution and cannot be evaluated at fixed points, so (5.4) might be meaningless for a general model. It turns out that the definition actually always makes sense, provided that the second identity in (5.4) is interpreted as

$$f_z(\mathcal{J}_k(\tau)) = \sum_{n\ge 0}\big(D^{(k)}K_n\star\Pi_z\tau\big)(z)\;,$$

where $K = \sum_{n\ge 0}K_n$ as in (2.1). This is because the bound (2.1), combined with the bound (4.5) and the fact that K_n is supported in the ball of radius 2^{-n} imply that

$$\big|\big(D^{(k)}K_n\star\Pi_z\tau\big)(z)\big| \lesssim 2^{(|k|-|\tau|-2)n}\;.$$

The condition $|k| < |\tau| + 2$ appearing in (5.3c) is then precisely what is required to guarantee that this is always summable.

5.5. Abstract fixed point problem. We now show how to reformulate a stochastic PDE as a fixed point problem in some space $\mathcal{D}^\gamma$ based on an admissible model for the regularity structure associated to the SPDE by the construction of Section 5.2. For definiteness, we focus on the example of the KPZ equation (1.3), but all other examples mentioned in the introduction can be treated in virtually the same way. Writing P for the heat kernel, the mild formulation of (1.3) is given by

$$h = P \star \mathbf{1}_{t>0}\big((\partial_x h)^2 + \xi\big) + Ph_0 \;, \tag{5.7}$$

where we write Ph_0 for the harmonic extension of h_0. (This is just the solution to the heat equation with initial condition h_0.) In order to formulate this as a fixed point problem in $\mathcal{D}^\gamma$ for a suitable value of $\gamma > 0$, we will make use of the following far-reaching extension of Schauder's theorem.

Theorem 5.9. *Fix one of the regularity structures built in the previous section and fix an admissible model. Then, for all but a discrete set of values of $\gamma > 0$, there exists a continuous operator $\mathcal{P} \colon \mathcal{D}^\gamma \to \mathcal{D}^{\gamma+2}$ such that the identity*

$$\mathcal{R}\mathcal{P}f = P \star \mathcal{R}f \;, \tag{5.8}$$

holds for every $f \in \mathcal{D}^\gamma$. Furthermore, one has $\big(\mathcal{P}f\big)(z) - \mathcal{I}f(z) \in \bar{T}$.

Remark 5.10. Recall that $\bar{T} \subset T$ denotes the linear span of the X^k, which represent the usual Taylor polynomials. Again, while $\mathcal{P}$ is a linear map when we consider the underlying model as fixed, it can (and should) also be viewed as a continuous nonlinear map from $\mathscr{M}_0 \ltimes \mathcal{D}^\gamma$ into $\mathscr{M}_0 \ltimes \mathcal{D}^{\gamma+2}$. The reason why some values of γ need to be excluded is essentially the same as for the usual Schauder theorem.

For a proof of Theorem 5.9 and a precise description of the operator $\mathcal{P}$, see [23, Sec. 5]. With the help of the operator $\mathcal{P}$, it is then possible to reformulate (5.7) as the following fixed point problem in $\mathcal{D}^\gamma$, provided that we have an admissible model at our disposal:

$$H = \mathcal{P}\mathbf{1}_{t>0}\big((\mathscr{D}H)^2 + \Xi\big) + Ph_0 \;. \tag{5.9}$$

Here, the smooth function Ph_0 is interpreted as an element in $\mathcal{D}^\gamma$ with values in $\bar{T}$ via its Taylor expansion of order γ. Note that in the context of the regularity structure associated to the KPZ equation in Section 5.2, the right hand side of this equation makes sense for every $H \in \mathcal{D}^\gamma$, provided that H takes values in $T_{\mathcal{U}}$. This is an immediate consequence of the property (5.1).

Remark 5.11. As already mentioned earlier, we cheat here in the sense that $\mathcal{D}^\gamma$ should really be replaced by a space $\mathcal{D}^{\gamma,\eta}$ allowing for a suitable singular behaviour on the hyperplane $t = 0$.

It is also possible to show (see [23, Thm 4.7]) that if we set $|\Xi| = -\frac{3}{2} - \kappa$ for some sufficiently small $\kappa > 0$, then one has $(\mathscr{D}H)^2 \in \mathcal{D}^{\gamma - \frac{3}{2} - \kappa}$ for $H \in \mathcal{D}^\gamma$. As a consequence, we expect to be able to find local solutions to the fixed point problem (5.9), provided that we

formulate it in $\mathcal{D}^\gamma$ for $\gamma > \frac{3}{2} + \kappa$. This is indeed the case, and a more general instance of this fact can be found in [23, Thm 7.8]. Furthermore, the local solution is locally Lipschitz continuous as a function of both the initial condition h_0 and the underlying admissible model $(\Pi, \Gamma) \in \mathscr{M}_0$.

Now that we have a local solution $H \in \mathcal{D}^\gamma$ for (5.9), we would like to know how this solution relates to the original problem (1.3). This is given by the following simple fact:

Proposition 5.12. *If the underlying model* (Π, Γ) *is built from a smooth function* ξ_ε *as in* (5.3) *and if* H *solves* (5.9)*, then* $\mathcal{R}H$ *solves* (5.7)*.*

Proof. As a consequence of (5.8), we see that $\mathcal{R}H$ solves

$$\mathcal{R}H = P \star \mathbf{1}_{t>0}\big(\mathcal{R}\big((\mathscr{D}H)^2\big) + \xi_\varepsilon\big) + Ph_0 \;.$$

Combining (5.3b) with (4.10), it is not difficult to see that in this particular case, one has $\mathcal{R}\big((\mathscr{D}H)^2\big) = (\partial_x \mathcal{R}H)^2$, so that the claim follows. □

The results of the previous subsection yield a robust solution theory for (5.9) which projects down (via $\mathcal{R}$) to the usual solution theory for (1.3) for smooth driving noise ξ_ε. If it were the case that the sequence of models $(\Pi^{(\varepsilon)}, \Gamma^{(\varepsilon)})$ associated to the regularised noise ξ_ε via (5.3) converges to a limit in $\mathscr{M}_0$, then this would essentially conclude our analysis of (1.3).

Unfortunately, this is *not* the case. Indeed, in all of the examples mentioned in the introduction except for (1.2), the sequence of models $(\Pi^{(\varepsilon)}, \Gamma^{(\varepsilon)})$ does not converge as $\varepsilon \to 0$. In order to remedy to this situation, the idea is to look for a sequence of "renormalised" models $(\hat\Pi^{(\varepsilon)}, \hat\Gamma^{(\varepsilon)})$ which are also admissible and also satisfy $\hat\Pi_z^{(\varepsilon)}\Xi = \xi_\varepsilon$, but do converge to a limit as $\varepsilon \to 0$. The last section of this article shows how these renormalised models can be constructed.

5.6. Renormalisation. In order to renormalise our model, we will build a very natural group of continuous transformations of $\mathscr{M}_0$ that build a new admissible model from an old one. The renormalised model will then be the image of the "canonical" model $(\Pi^{(\varepsilon)}, \Gamma^{(\varepsilon)})$ under a (diverging) sequence of such transformations. Since we want the new model to also be admissible, the only defining property that we are allowed to modify in (5.3) is the definition of the product. In order to describe the renormalised model, it turns out to be more convenient to consider again its representation by a single linear map $\hat{\mathbf{\Pi}}^{(\varepsilon)} : T \to \mathcal{S}'$ as in (5.3), which is something we can do by Remark 5.8.

At this stage, we do not appear to have much choice: the only "reasonable" way of building $\hat{\mathbf{\Pi}}^{(\varepsilon)}$ from $\mathbf{\Pi}^{(\varepsilon)}$ is to compose it to the right with some fixed linear map $M_\varepsilon : T \to T$:

$$\hat{\mathbf{\Pi}}^{(\varepsilon)} = \mathbf{\Pi}^{(\varepsilon)} M_\varepsilon \;. \tag{5.10}$$

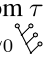
If we do this for an arbitrary map M_ε, we will of course immediately lose the algebraic and analytical properties that allow to associate an admissible model $(\hat\Pi^{(\varepsilon)}, \hat\Gamma^{(\varepsilon)})$ to the map $\hat{\mathbf{\Pi}}^{(\varepsilon)}$. As a matter of fact, it is completely unclear *a priori* whether there exists *any* non-trivial map M_ε that preserves these properties. Fortunately, these maps do exists and a somewhat indirect characterisation of them can be found in [23, Sec. 8]. Even better, there are sufficiently many of them so that the divergencies of $\mathbf{\Pi}^{(\varepsilon)}$ can be compensated by a judicious choice of M_ε.

Let us just illustrate how this plays out in the case of the KPZ equation already studied in the last subsection. In order to simplify notations, we now use the following shorthand graphical notation for elements of $\mathcal{U} \cup \mathcal{V}$. For Ξ, we draw a small circle. The integration map $\mathcal{I}$ is then represented by a downfacing wavy line and $\mathscr{D}\mathcal{I}$ is represented by a downfacing plain line. The multiplication of symbols is obtained by joining them at the root. For example, we have

$$(\mathscr{D}\mathcal{I}(\Xi))^2 =$$  , $$(\mathscr{D}\mathcal{I}(\mathscr{D}\mathcal{I}(\Xi)^2))^2 =$$, $$\mathcal{I}(\mathscr{D}\mathcal{I}(\Xi)^2) =$$  .

In the case of the KPZ equation, it turns out that one can exhibit an explicit four-parameter group of matrices M which preserve admissible models when used in (5.10). These matrices are of the form $M = \exp(-\sum_{i=0}^{3} C_i L_i)$, where the generators L_i are determined by the following contraction rules:

L_0 :  $\mapsto \mathbf{1}$, L_1 :  $\mapsto \mathbf{1}$, L_2 :  $\mapsto \mathbf{1}$ L_3 :  $\mapsto \mathbf{1}$. (5.11)

This should be understood in the sense that if τ is an arbitrary formal expression, then $L_0\tau$ is the sum of all formal expressions obtained from τ by performing a substitution of the type  $\mapsto \mathbf{1}$. For example, one has L_0  $= 2$  , L_0  $= 2$  $+$  , etc. The extension of the other operators L_i to all of T is given by $L_i\tau = 0$ for $i \neq 0$ and every τ for which L_i wasn't already defined in (5.11). We then have the following result, which is a consequence of [23, Sec. 8] and [28] and was implicit in [21]:

Theorem 5.13. *Let M_ε be given as above, let $\mathbf{\Pi}^{(\varepsilon)}$ be constructed from ξ_ε as in* (5.2)*, and let $\hat{\mathbf{\Pi}}^{(\varepsilon)} = \mathbf{\Pi}^{(\varepsilon)} M_\varepsilon$. Then, there exists a unique admissible model $(\hat\Pi^{(\varepsilon)}, \hat\Gamma^{(\varepsilon)})$ such that $\hat\Pi_z^{(\varepsilon)} = \hat{\mathbf{\Pi}}^{(\varepsilon)} \hat F_z^{(\varepsilon)}$, where $\hat F_z^{(\varepsilon)}$ relates to $\hat\Pi_z^{(\varepsilon)}$ as in* (5.4). *Furthermore, one has the identity*

$$\big(\hat\Pi_z^{(\varepsilon)}\tau\big)(z) = \big(\Pi_z^{(\varepsilon)} M_\varepsilon \tau\big)(z) \;. \tag{5.12}$$

Finally, there is a choice of M_ε such that $(\hat\Pi^{(\varepsilon)}, \hat\Gamma^{(\varepsilon)})$ converges to a limit $(\hat\Pi, \hat\Gamma)$ which is universal in that it does not depend on the details of the regularisation procedure.

Remark 5.14. Despite (5.12), it is not true in general that $\hat\Pi_z^{(\varepsilon)} = \Pi_z^{(\varepsilon)} M_\varepsilon$. The point is that (5.12) only holds at the point z and not at $z' \neq z$.

In order to complete our survey of Theorem 1.1, it remains to identify the solution to (5.9) with respect to the renormalised model $(\hat\Pi^{(\varepsilon)}, \hat\Gamma^{(\varepsilon)})$ with the classical solution to some modified partial differential equation. The continuity of the abstract solution map then immediately implies that the solutions to the modified PDE converge to a limit. The fact that the limiting model $(\hat\Pi, \hat\Gamma)$ is universal also implies that this limit is universal.

Theorem 5.15. *Let $M_\varepsilon = \exp(-\sum_{i=0}^{3} C_i^{(\varepsilon)} L_i)$ be as above and let $(\hat\Pi^{(\varepsilon)}, \hat\Gamma^{(\varepsilon)})$ be the corresponding renormalised model. Let furthermore H be the solution to* (5.9) *with respect to this model. Then, the function $h(t,x) = \big(\mathcal{R}H\big)(t,x)$ solves the equation*

$$\partial_t h = \partial_x^2 h + (\partial_x h)^2 - 4C_0^{(\varepsilon)}\, \partial_x h + \xi_\varepsilon - (C_1^{(\varepsilon)} + C_2^{(\varepsilon)} + 4C_3^{(\varepsilon)}) \;. \tag{5.13}$$

Remark 5.16. In order to obtain a limit $(\hat\Pi, \hat\Gamma)$, the renormalisation constants $C_i^{(\varepsilon)}$ should be chosen in the following way:

$$C_0^{(\varepsilon)} = 0 \;, \qquad C_1^{(\varepsilon)} = \frac{c_1}{\varepsilon} \;, \qquad C_2^{(\varepsilon)} = 4c_2 \log\varepsilon + c_3 \;, \qquad C_3^{(\varepsilon)} = -c_2 \log\varepsilon + c_4 \;.$$

Here, the c_i are constants of order 1 that depend on the details of the regularisation procedure for ξ_ε. The fact that $C_0^{(\varepsilon)} = 0$ explains why the corresponding term does not appear in (1.3). The fact that the diverging parts of $C_2^{(\varepsilon)}$ and $C_3^{(\varepsilon)}$ cancel in (5.13) explains why this logarithmic sub-divergence was not observed in [4] for example.

Proof. We first note that, as a consequence of Theorem 5.9 and of (5.9), one can write for $t > 0$

$$H = \mathcal{I}\big((\mathscr{D}H)^2 + \Xi\big) + (...)\,, \tag{5.14}$$

where $(...)$ denotes some terms belonging to $\bar{T} \subset T$.

By repeatedly using this identity, we conclude that any solution $H \in \mathcal{D}^\gamma$ to (5.9) for γ greater than (but close enough to) $3/2$ is necessarily of the form

$$H = h\,\mathbf{1} + \mathcal{I}(\xi) + \mathcal{I}(\mathcal{I}(\xi)\mathcal{I}(\xi)) + h'\,X_1 + 2\mathcal{I}(\mathcal{I}(\xi)X_1) + 2h'\,\mathcal{I}(X_1^2)\,, \tag{5.15}$$

for some real-valued functions h and h'. Note that h' is treated as an independent function here, we certainly do not mean to suggest that the function h is differentiable! Our notation is only by analogy with the classical Taylor expansion. As an immediate consequence, $\mathscr{D}H$ is given by

$$\mathscr{D}H = \mathcal{I}(\xi) + \mathcal{I}(\mathcal{I}(\xi)\mathcal{I}(\xi)) + h'\,\mathbf{1} + 2\mathcal{I}(\mathcal{I}(\xi)X_1) + 2h'\,\mathcal{I}(X_1^2)\,, \tag{5.16}$$

as an element of $\mathcal{D}^\gamma$ for γ close to $1/2$. The right hand side of the equation is then given up to order 0 by

$$(\mathscr{D}H)^2 + \Xi = \Xi + \mathcal{I}(\xi\mathcal{I}(\xi)\mathcal{I}(\xi)) + 2\mathcal{I}(\mathcal{I}(\xi)\mathcal{I}(\xi)X_1) + 2h'\,\mathcal{I}(\xi) + \mathcal{I}(\xi\mathcal{I}(\xi)\mathcal{I}(\xi)\mathcal{I}(\xi)) + 4\mathcal{I}(\mathcal{I}(\xi)\mathcal{I}(\xi)X_1) + 2h'\,\mathcal{I}(\mathcal{I}(\xi)X_1) + 4h'\,\mathcal{I}(\mathcal{I}(\xi)X_1) + (h')^2\,\mathbf{1}\,. \tag{5.17}$$

Using the definition of M_ε, we conclude that

$$M_\varepsilon \mathscr{D}H = \mathscr{D}H - 4C_0^{(\varepsilon)}\,\mathcal{I}(X_1^2)\,,$$

so that, as an element of $\mathcal{D}^\gamma$ with very small (but positive) γ, one has the identity

$$(M_\varepsilon \mathscr{D}H)^2 = (\mathscr{D}H)^2 - 8C_0^{(\varepsilon)}\mathcal{I}(\mathcal{I}(\xi)X_1)\,.$$

As a consequence, after neglecting again all terms of strictly positive homogeneity, one has the identity

$$M_\varepsilon\big((\mathscr{D}H)^2 + \Xi\big) = (M_\varepsilon \mathscr{D}H)^2 + \Xi - 4C_0^{(\varepsilon)}\,M_\varepsilon \mathscr{D}H - (C_1^{(\varepsilon)} + C_2^{(\varepsilon)} + 4C_3^{(\varepsilon)})\,.$$

Combining this with (5.12) and (4.10), we conclude that

$$\mathcal{R}\big((\mathscr{D}H)^2 + \Xi\big) = (\partial_x \mathcal{R}H)^2 + \xi_\varepsilon - 4C_0^{(\varepsilon)}\,\partial_x \mathcal{R}H - (C_1^{(\varepsilon)} + C_2^{(\varepsilon)} + 4C_3^{(\varepsilon)})\,,$$

from which the claim then follows in the same way as for Proposition 5.12. □

Remark 5.17. Ultimately, the reason why the theory mentioned in Section 1.1 (or indeed the theory of controlled rough paths, as originally exploited in [21]) can also be applied in this case is that in (5.15), only *one* basis vector besides those in $\mathcal{T}$ (i.e. besides $\mathbf{1}$ and X_1) comes with a non-constant coefficient, namely the basis vector $\mathcal{I}(X_1^2)$. The methodology explained in Section 3.1 on the other hand can be applied whenever no basis vector besides those in $\mathcal{T}$ comes with a non-constant coefficient.

Acknowledgements. I am delighted to thank the Institute for Advanced Study for its warm hospitality and the 'The Fund for Math' for funding my stay there. This work was supported by the Leverhulme trust through a leadership award, the Royal Society through a Wolfson research award, and the ERC through a consolidator award.

References

[1] S. Albeverio and M. Röckner, *Stochastic differential equations in infinite dimensions: solutions via Dirichlet forms*, Probab. Theory Related Fields **89**(3) (1991), 347–386.

[2] H. Bahouri, J.-Y. Chemin, and R. Danchin, *Fourier analysis and nonlinear partial differential equations*, volume 343 of Grundlehren der Mathematischen Wissenschaften, Springer, Heidelberg, 2011.

[3] Á. Bényi, D. Maldonado, and V. Naibo, *What is ... a paraproduct?*, Notices Amer. Math. Soc. **57**(7) (2010), 858–860.

[4] L. Bertini and G. Giacomin, *Stochastic Burgers and KPZ equations from particle systems*, Comm. Math. Phys. **183**(3) (1997), 571–607.

[5] L. Bertini, E. Presutti, B. Rüdiger, and E. Saada, *Dynamical fluctuations at the critical point: convergence to a nonlinear stochastic PDE*, Teor. Veroyatnost. i Primenen. **38**(4) (1993), 689–741.

[6] J.-M. Bismut, *Martingales, the Malliavin calculus and hypoellipticity under general Hörmander's conditions*, Z. Wahrsch. Verw. Gebiete **56**(4) (1981), 469–505.

[7] J.-M. Bony, *Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires*, Ann. Sci. École Norm. Sup. (4) **14**(2) (1981), 209–246.

[8] R. A. Carmona and S. A. Molchanov *Parabolic Anderson problem and intermittency*, Mem. Amer. Math. Soc. **108**(518) (1994), viii+125.

[9] R. Catellier and K. Chouk, *Paracontrolled distributions and the 3-dimensional stochastic quantization equation*, ArXiv e-prints, Oct. 2013.

[10] G. Da Prato and A. Debussche, *Two-dimensional Navier-Stokes equations driven by a space-time white noise*, J. Funct. Anal. **196**(1) (2002), 180–210.

[11] ———, *Strong solutions to the stochastic quantization equations*, Ann. Probab. **31**(4) (2003), 1900–1916.

[12] ———, *A modified Kardar-Parisi-Zhang model*, Electron. Comm. Probab. **12** (2007), 442–453 (electronic).

[13] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, volume 44 of Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1992.

[14] R. L. Dobrushin, *Gaussian and their subordinated self-similar random generalized fields*, Ann. Probab. **7**(1) (1979), 1–28.

[15] P. K. Friz and M. Hairer, *A course on rough paths*, Universitext, Springer, 2014. To appear.

[16] P. K. Friz and N. B. Victoir, *Multidimensional stochastic processes as rough paths*, volume 120 of Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 2010. Theory and applications.

[17] P. Goncalves and M. Jara, *Universality of KPZ equation*, ArXiv e-prints, Mar. 2010.

[18] M. Gubinelli, *Controlling rough paths*, J. Funct. Anal. **216**(1) (2004), 86–140.

[19] M. Gubinelli, P. Imkeller, and N. Perkowski, *Paraproducts, rough paths and controlled distributions*, ArXiv e-prints, Oct. 2012.

[20] M. Hairer, *Rough stochastic PDEs*, Comm. Pure Appl. Math. **64**(11) (2011), 1547–1585.

[21] ———, *Solving the KPZ equation*, Ann. of Math. (2) **178**(2) (2013), 559–664.

[22] ———, *Introduction to regularity structures*, ArXiv e-prints, Jan. 2014. Braz. J. Prob. Stat., to appear.

[23] ———, *A theory of regularity structures*, Invent. Math., Mar. 2014.

[24] M. Hairer and J. Maas, *A spatial version of the Itô-Stratonovich correction*, Ann. Probab. **40**(4) (2012), 1675–1714.

[25] M. Hairer, J. Maas, and H. Weber, *Approximating rough stochastic PDEs*, Comm. Pure Appl. Math. **67**(5) (2014), 776–870.

[26] M. Hairer, É. Pardoux, and A. Piatnitksy, *A Wong-Zakai theorem for stochastic PDEs*, Work in progress, 2014.

[27] M. Hairer and N. S. Pillai, *Regularity of laws and ergodicity of hypoelliptic SDEs driven by rough paths*, Ann. Probab. **41**(4) (2013), 2544–2598.

[28] M. Hairer and J. Quastel, *Continuous interface models rescale to KPZ*, Work in progress, 2014.

[29] M. Hairer and H. Weber, *Rough Burgers-like equations with multiplicative noise*, Probab. Theory Related Fields **155**(1-2) (2013), 71–126.

[30] K. Itô, *Stochastic integral*, Proc. Imp. Acad. Tokyo **20** (1944), 519–524.

[31] G. Jona-Lasinio and P. K. Mitter, *On the stochastic quantization of field theory*, Comm. Math. Phys. **101**(3) (1985), 409–436.

[32] M. Kardar, G. Parisi, and Y.-C. Zhang, *Dynamic scaling of growing interfaces*, Phys. Rev. Lett., **56**(9) (Mar. 1986), 889–892.

[33] S. Kusuoka and D. Stroock, *Applications of the Malliavin calculus. I*, In Stochastic analysis (Katata/Kyoto, 1982), volume 32 of North-Holland Math. Library, pp. 271–306. North-Holland, Amsterdam, 1984.

[34] T. J. Lyons, *Differential equations driven by rough signals*, Rev. Mat. Iberoamericana **14**(2) (1998), 215–310.

[35] T. J. Lyons, M. Caruana, and T. Lévy, *Differential equations driven by rough paths*, volume 1908 of Lecture Notes in Mathematics, Springer, Berlin, 2007. Lectures from the 34th Summer School on Probability Theory held in Saint-Flour.

[36] T. J. Lyons and Z. Qian, *System control and rough paths*, Oxford Mathematical Monographs. Oxford University Press, Oxford, 2002. Oxford Science Publications.

[37] P. Malliavin, *Stochastic analysis*, volume 313 of Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, Berlin, 1997.

[38] J. Norris, *Simplified Malliavin calculus*, In Séminaire de Probabilités, XX, 1984/85, volume 1204 of Lecture Notes in Math., pp. 101–130. Springer, Berlin, 1986.

[39] D. Nualart, *The Malliavin calculus and related topics*, Probability and its Applications (New York). Springer-Verlag, Berlin, second edition, 2006.

[40] G. Parisi and Y. S. Wu, *Perturbation theory without gauge fixing*, Sci. Sinica **24**(4) (1981), 483–496.

[41] J. Schauder, *Über lineare elliptische Differentialgleichungen zweiter Ordnung*, Math. Z. **38**(1) (1934), 257–282.

[42] L. Simon, *Schauder estimates by scaling*, Calc. Var. Partial Differential Equations **5**(5) (1997), 391–407.

[43] R. L. Stratonovič, *A new form of representing stochastic integrals and equations*, Vestnik Moskov. Univ. Ser. I Mat. Meh. **1964**(1) (1964), 3–12.

[44] E. Wong and M. Zakai, *On the relation between ordinary and stochastic differential equations*, Internat. J. Engrg. Sci. **3** (1965), 213–229.

Mathematics Institute, The University of Warwick, U.K.
E-mail: M.Hairer@Warwick.ac.uk

Anomalous random walks and diffusions: From fractals to random media

Takashi Kumagai

Abstract. We present results concerning the behavior of random walks and diffusions on disordered media. Examples treated include fractals and various models of random graphs, such as percolation clusters, trees generated by branching processes, Erdős-Rényi random graphs and uniform spanning trees. As a consequence of the inhomogeneity of the underlying spaces, we observe anomalous behavior of the corresponding random walks and diffusions. In this regard, our main interests are in estimating the long time behavior of the heat kernel and in obtaining a scaling limit of the random walk. We will overview the research in these areas chronologically, and describe how the techniques have developed from those introduced for exactly self-similar fractals to the more robust arguments required for random graphs.

Mathematics Subject Classification (2010). Primary 60J45; Secondary 05C81, 60K37.

Keywords. Fractals, heat kernel estimates, percolation, random media, sub-diffusivity.

1. Introduction

Since the mid-sixties, mathematical physicists have investigated anomalous behavior of random walks and diffusions on disordered media (see for example [17]). The random walk on a percolation cluster – the so-called 'ant in the labyrinth' ([24]) – is one of the central examples. Recall that the bond percolation model on the lattice $\mathbb{Z}^d$, $d \geq 2$, is defined as follows: each nearest neighbor bond is open with probability $p \in [0,1]$ and closed otherwise, independently of all the others. It is well-known that this model exhibits a phase transition, whereby if $\theta(p) := P_p(|\mathcal{C}(0)| = +\infty)$, where $\mathcal{C}(0)$ is the open cluster containing 0, then there exists $p_c = p_c(\mathbb{Z}^d) \in (0,1)$ such that $\theta(p) = 0$ if $p < p_c$ and $\theta(p) > 0$ if $p > p_c$. For $p > p_c$, there exists a unique open infinite cluster upon which the long time behavior of the simple random walk is similar to that of the simple random walk on $\mathbb{Z}^d$ (see Section 4.1). For the simple random walk on the critical percolation cluster, however, in 1982 Alexander and Orbach [1] made a striking conjecture about how there might be quite different behavior. (To make the problem mathematically precise, one has to consider the critical percolation cluster conditioned to be infinite, as we discuss in Section 4.2.) Let $Y = \{Y_n^\omega\}_{n\in\mathbb{N}}$ be the simple random walk on the cluster (i.e. Y_n^ω is in one of the adjacent neighbors of Y_{n-1}^ω with equal probabilities), and $p_n^\omega(x,y)$ be its heat kernel (transition density); see (3.3) for precise definition. Here and in the following, the suffix ω stands for the randomness of the media.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Define

$$d_s := -2 \lim_{n\to\infty} \frac{\log p^{\omega}_{2n}(x,x)}{\log n} \tag{1.1}$$

as the spectral dimension of the cluster if the limit exists. (To be precise, the original definition of d_s was the 'density of states', which gives the asymptotic growth of the eigenvalue counting function.) One formulation of the Alexander-Orbach conjecture is that $d_s = 4/3$ for all $d \geq 2$. Clearly, this expresses anomalous behavior for the random walk, since $d_s = d$ for simple random walk on $\mathbb{Z}^d$. These works stimulated a lot of interest from mathematical physicists in exact fractals as well (see for example [41]).

Mathematical progress on these problems started to be made in the late eighties. In 1986, Kesten wrote two beautiful papers ([31, 32]) in which he constructed an 'incipient infinite cluster' for critical percolation on $\mathbb{Z}^2$ and showed that the random walk on this was anomalous (in the latter work, he also considered random walks on critical models of trees); these were the first significant mathematically rigorous works in this area. Kesten's work and mathematical physicists' work mentioned above triggered intensive research on diffusions on fractals, which are "ideal" disordered media. As part of this, Brownian motion was constructed on typical fractals, such as the Sierpinski gasket, and properties of these processes were obtained (see Section 2). These included detailed heat kernel estimates of the so-called sub-Gaussian form, meaning that the heat kernel is bounded from above and below by

$$c_1 t^{-d_s/2} \exp\Big(- c_2 \Big(\frac{d(x,y)^{d_w}}{t} \Big)^{1/(d_w-1)} \Big)$$

with different pairs of constants (c_1, c_2) for the upper and lower bounds. Here $d_w > 2$ is a constant and $d(\cdot,\cdot)$ is a geodesic distance on the fractal.

While diffusions on fractals had been extensively studied by 2000 and continue to be actively studied, the turn of the century saw increasing moves being made to analyze "fractal-like spaces" instead of working only on ideal fractals. The key issue here is whether the sub-Gaussian estimates mentioned above are stable under perturbations of spaces and operators. (Note that when $d_s = d$ and $d_w = 2$, the corresponding estimates are Gaussian estimates, and such a perturbation theory was extensively developed in the nineties.) In this direction, several functional inequalities have been shown to be equivalent to the sub-Gaussian estimates, some of which are stable under perturbations, meaning that the stability problem has been affirmatively resolved (see Section 3).

It turns out that such a stability theory is useful even for the analysis on random media, including percolation clusters as Kesten considered. Indeed, some functional inequalities have been modified and applied to random walks on various models of disordered media, especially on percolation clusters (see Section 4). Specifically, the Alexander-Orbach conjecture has been affirmatively solved for high dimensions (Theorem 4.4). For some models, scaling limits of random walks have also been established (see Section 4.1 and Section 5); these include supercritical percolation clusters, critical branching processes conditioned to be large, the Erdős-Rényi random graph in the critical window, and the 2-dimensional uniform spanning tree.

The aim of this paper is to give a overview of the stream of research introduced above. It is a very restricted survey and the references are far from complete. Due to space restriction, for papers which are very important but for which details are not discussed in this paper, names of authors and years of publication are mentioned but without inclusion in the list of

references. We apologize to the authors of relevant papers which are not cited here. Readers can find more detailed information in the following books/surveys [5, 7, 17, 19, 23, 25, 27, 29, 33, 34, 36, 38, 39, 42, 44, 45].

Notation. We write $f \asymp g$ if there exist constants $c_1, c_2 > 0$ such that $c_1 g(x) \le f(x) \le c_2 g(x)$ for all x, and $f \sim g$ if $\lim_{|x|\to\infty} f(x)/g(x) = 1$.

2. Anomalous heat transfer on fractals

Let $a = (0,0), b = (1,0), c = (1/2, \sqrt{3}/2)$, and set

$$F_1(x) = (x-a)/2 + a,\ F_2(x) = (x-b)/2 + b \text{ and } F_3(x) = (x-c)/2 + c.$$

Then, there exists unique non-void compact set such that $K = \cup_{i=1}^3 F_i(K)$; we call K the 2-dimensional Sierpinski gasket. Define the unbounded Sierpinski gasket as $\hat{K} = \cup_{n=0}^{\infty} 2^n K$.

We first explain the construction of Brownian motion on $\hat{K}$. Let

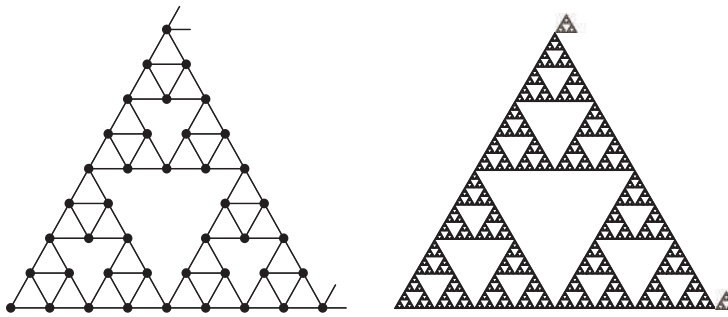


Figure 2.1. Sierpinski gasket graph V_0 and Sierpinski gasket $\hat{K}$

$$V_0 = \bigcup_{m=0}^{\infty} 2^m \Big(\bigcup_{i_1,\cdots,i_m=1}^{3} F_{i_1} \circ \cdots \circ F_{i_m}(\{a,b,c\}) \Big),\ \ V_m = 2^{-m} V_0.$$

The closure of $\cup_{m\ge 0} V_m$ is $\hat{K}$. Let $\{X(i)\}_{i\ge 0}$ be the simple random walk on V_0. That is, it is a random walk such that $X(i+1)$ is in one of the adjacent neighbors of $X(i)$ in V_0 (i.e. points in the same triangles with length 1 as those $X(i)$ belongs to) with equal probabilities. Let $X_m(i) := 2^{-m} X(i)$ be the simple random walk on V_m. Since X_m moves distance 2^{-m} per unit time, $X_m(i) \to 0$ as $m \to \infty$ for fixed i. So, we must speed up the random walks in order to obtain a non-trivial limit. It is plausible to choose the time scale as the average time for the random walk on V_{m+1} starting from a point in V_m to reach one of the neighboring points in V_m. By the self-similarity and symmetry of $\hat{K}$, this average time is independent of m and it is equal to the average time for X_1 starting from a to arrive at either b or c. A simple calculation deduces that the value is 5. Let $Y_t^{(m)} := X_m([5^m t])$. Then, it can be proved that $\{Y^{(m)}\}$ converges to a non-trivial diffusion on $\hat{K}$ as $m \to \infty$, which is called Brownian motion on $\hat{K}$. (One can construct Brownian motion on K similarly.) Brownian motion on the gasket was first constructed by Goldstein (1987) and Kusuoka (1987) independently.

Characterization of Brownian motion is also known; any self-similar diffusion process on $\hat{K}$ whose law is invariant under local translations and reflections on each small triangle is a constant time change of this diffusion ([16]).

The corresponding Laplacian Δ is defined as follows:

$$\Delta f(x) = \lim_{m\to\infty} 5^m \Big(\sum_{x_i:\, x \overset{m}{\sim} x_i} f(x_i) - 4f(x) \Big), \quad x \in \cup_{m\geq 0} V_m \setminus \{0\},$$

for f in a suitable function space, where $x \overset{m}{\sim} y$ means that x and y are adjacent in V_m. Note that the standard approximation for the Laplacian on $\mathbb{R}$ is $\Delta f(x) = \lim_{m\to\infty} 2^{2m}(f(x + 2^{-m}) + f(x - 2^{-m}) - 2f(x))$ for $f \in C^2(\mathbb{R})$. Set $d_w = \log 5/\log 2$ so that $5 = 2^{d_w}$. Naively, we can say that the Laplacian on the gasket is a "differential operator of order d_w". (One way of stating this rigorously is that the domain of the corresponding Dirichlet form on the gasket is a Besov space of order $d_w/2$ (Jonsson (1996), Grigor'yan-Hu-Lau (2003)).) Kigami (1989) was the first to construct the Laplacian on the gasket directly. It turns out that the theory of Dirichlet forms ([23]) is well-applicable to this area, and diffusions (self-adjoint operators) on fractals have been constructed through Dirichlet forms systematically. Fukushima-Shima (1992) is one of the first who applied the Dirichlet form theory to fractals.

On $\mathbb{R}^d$, we can define $\hat{K}$ similarly from the family of $(d+1)$-th contraction maps with contraction rate $1/2$. (For $d = 1$, $\hat{K} = [0, \infty)$.) The Hausdorff dimension of the d-dimensional gasket is $d_f = \log(d+1)/\log 2$. The time scaling is $d + 3$ and $d_w = \log(d+3)/\log 2$.

In order to understand the asymptotic properties of the process, it is very important and useful to obtain detailed heat kernel estimates. Let $\{B(t)\}_{t\geq 0}$ be Brownian motion on the gasket and define

$$P_t f(x) = E^x[f(B(t))] = \int_{\hat{K}} p_t(x,y) f(y) \mu(dy),$$

where μ is the normalized Hausdorff measure on $\hat{K}$. $\{P_t\}_{t\geq 0}$ is the semigroup and $p_t(\cdot,\cdot)$ is the heat kernel (transition density) for Brownian motion on $\hat{K}$. $p_t(\cdot,\cdot)$ is a fundamental solution of the heat equation for the Laplacian. For the case of Brownian motion on $\mathbb{R}^d$, $p_t(x,y)$ is the Gauss-kernel $\frac{1}{(2\pi t)^{d/2}} \exp(-|x-y|^2/(2t))$.

Let $d(x,y)$ be the shortest distance between x and y in $\hat{K}$. The following sub-Gaussian heat kernel estimates are obtained by Barlow-Perkins [16].

Theorem 2.1. *$p_t(x,y)$ obeys the following estimates for $t > 0, x, y \in \hat{K}$:*

$$\begin{aligned} c_1 t^{-d_f/d_w} \exp\Big(- c_2 \Big(\frac{d(x,y)^{d_w}}{t} \Big)^{1/(d_w-1)} \Big) &\leq p_t(x,y) \\ &\leq c_3 t^{-d_f/d_w} \exp\Big(- c_4 \Big(\frac{d(x,y)^{d_w}}{t} \Big)^{1/(d_w-1)} \Big). \end{aligned} \tag{2.1}$$

The simple random walk on V_0 also obeys (2.1) for $d(x,y) \leq t \in \mathbb{N}$ (Jones (1996)).

From the probabilistic viewpoint, d_w is the order of the diffusion speed of particles and it is called the walk dimension. Indeed, by integrating (2.1), we have $c_5 t^{1/d_w} \leq E^x[d(x, B(t))] \leq c_6 t^{1/d_w}$. As $d_w > 2$, the behavior of the process is anomalous (for a long time, it diffuses slower than Brownian motion on $\mathbb{R}^d$, so the behavior is sub-diffusive).

This diffusion does not have finite quadratic variation, so it is not a semi-martingale ([16]). Its martingale dimension is 1 (Kusuoka (1989), Hino (2008)). Set $d_s/2 = d_f/d_w$. This d_s, which is the same exponent as in (1.1), gives the asymptotic growth of the eigenvalue counting function for the Laplacian on K, and it is called the spectral dimension. Spectral properties of the Laplacian have been extensively studied (Fukushima-Shima (1992), Kigami-Lapidus (1993), Barlow-Kigami (1997), Teplyaev (1998), etc.). Unlike the Euclidean case, Brownian motion and the Laplacian on the gasket exhibit oscillations in their asymptotics; in the asymptotics of the eigenvalue counting function (Barlow-Kigami (1997)), in the on-diagonal heat kernel asymptotics (Grabner-Woess (1997), Kajino (2013)), and in Schilder's large-deviation principle (Ben Arous-Kumagai (2000)).

(2.1) is a very useful estimate. Various properties of Brownian motion such as laws of the iterated logarithm can be deduced from this estimate. It also implies nice regularity properties of caloric functions $u(t,x)$ (i.e. solutions of the heat equation $\frac{\partial u}{\partial t} = \Delta u$). For $S, R \in (0,\infty), x_0 \in \hat{K}$, set

$$Q_- = (S + R^{d_w}, S + 2R^{d_w}) \times B(x_0, R),\ Q_+ = (S + 3R^{d_w}, S + 4R^{d_w}) \times B(x_0, R).$$

The parabolic Harnack inequalities compare the values of caloric functions on Q_- and Q_+ uniformly. They imply uniform Hölder continuity of the caloric functions.

Theorem 2.2 (Generalized parabolic Harnack inequalities and Hölder continuity). *There exist $c_1, c_2, \theta > 0$ such that, for any $S, R \in (0,\infty), x_0 \in \hat{K}$, if u is a non-negative caloric function on $(S, S + 4R^{d_w}) \times B(x_0, 2R)$, then the following hold:*

$$\sup_{(t,x)\in Q_-} u(t,x) \le c_1 \inf_{(t,x)\in Q_+} u(t,x), \qquad (PHI(d_w))$$

$$|u(s,x) - u(s',x')| \le c_2 \left(\frac{|s-s'|^{1/d_w} + d(x,x')}{R} \right)^{\theta} \|u\|_\infty, \qquad (2.2)$$

for any $(s,x), (s',x') \in (S + R^{d_w}, S + 4R^{d_w}) \times B(x_0, R)$.

In fact, (2.1) and (PHI(d_w)) are equivalent under a suitable volume growth condition as we will see in the next section. (PHI(d_w)) implies various regularity properties of harmonic functions such as the elliptic Harnack inequalities and the Liouville property (i.e. if u is a non-negative harmonic function on $\hat{K}$, then u is a constant function).

For more general fractals such as nested fractals introduced by Lindstrøm (1990) and Sierpinski carpets (see Figure 2.2, the left figure is an example of nested fractals), Brownian motion is constructed and it is known that the heat kernels obey the sub-Gaussian estimates (2.1) (Barlow-Bass (1989, 1999), Lindstrøm (1990), Kumagai (1993), Fitzsimmons-Hambly-Kumagai (1994)). Characterization of Brownian motion on the fractals are also known (Metz (1996), Sabot (1997), Barlow-Bass-Kumagai-Teplyaev (2010)).

Open problem I. The existing construction of Brownian motion on the carpet requires detailed uniform control of harmonic functions (such as uniform Harnack inequalities) for the approximating processes; see for example [7]. Construct Brownian motion on the carpet without such detailed information.

We refer to [5, 7, 33, 34, 38, 44] for details on diffusions/analysis on fractals.

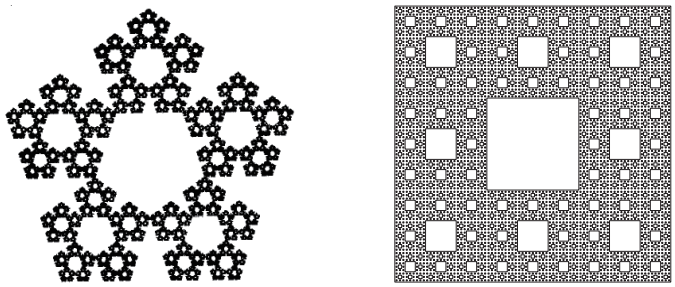


Figure 2.2. Penta-kun and Sierpinski carpet

3. Stability of parabolic Harnack inequalities and sub-Gaussian heat kernel estimates

Since fractals are "ideal" objects in that they have exact self-similarity, it is natural to ask if the inequalities (2.1) and (PHI(d_w)) are stable under perturbations of the state space and the operator.

Let us first briefly overview the history for the case of $d_w = 2$. For any divergence operator $\mathcal{L} = \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}(a_{ij}(x)\frac{\partial}{\partial x_j})$ on $\mathbb{R}^d$ satisfying a uniform elliptic condition, Aronson (1967) proved (2.1) with $d_f = d$ and $d_w = 2$. Later in the last century, there are outstanding results from the field of global analysis on manifolds. Let Δ be the Laplace-Beltrami operator on a complete Riemannian manifold M with the Riemannian metric $d(\cdot,\cdot)$ and with the Riemannian measure μ. Li-Yau (1986) proved the remarkable fact that if M has non-negative Ricci curvature, then the heat kernel $p_t(x,y)$ satisfies

$$c_1\Phi(x, c_2 d(x,y), t) \le p_t(x,y) \le c_3\Phi(x, c_4 d(x,y), t), \tag{3.1}$$

where $\Phi(x,r,t) = \mu(B(x,t^{1/2}))^{-1}\exp(-r^2/t)$. A few years later, Grigor'yan (1991) and Saloff-Coste (1992) refined the result and proved, in conjunction with the results by Fabes-Stroock (1986) and Kusuoka-Stroock (1987), that (3.1) is equivalent to a volume doubling condition (VD) plus Poincaré inequalities (PI(2)) –see Definition 3.1 and 3.3 for definitions in the graph setting. Their results were later extended to the framework of Dirichlet forms by Sturm (1996) and graphs by Delmotte (1999). Detailed heat kernel estimates are strongly related to the control of harmonic functions. The origin of ideas and techniques used in this field goes back to De Giorgi (1957), Nash (1958), Moser (1961,1964) and there are many other significant works in this area. See for example [25, 42] and the references therein. Summarizing, the following equivalence holds:

$$(3.1) \Leftrightarrow (\mathrm{VD}) + (\mathrm{PI}(2)) \Leftrightarrow (\mathrm{PHI}(2)). \tag{3.2}$$

Since (VD) and (PI(2)) are stable under some perturbations, we see that (3.1) and (PHI(2)) are also stable under the perturbations.

We will discuss the extension of (3.2) to the $d_w > 2$ case. Though such a generalization has also been established under a metric measure space with a local regular Dirichlet form, for simplicity, we will restrict our attention to the graph setting. We first set up notation and definitions.

3.1. Setting. Let G be a countably infinite set, and E a subset of $\{\{x,y\} \in G \times G : x \neq y\}$. We write $x \sim y$ if $\{x,y\} \in E$. A graph is a pair (G,E) and the graph distance $d(x,y)$ for $x,y \in G$ is the length of the shortest path from x to y (we set $d(x,x) = 0$). We assume the graph is connected (i.e. $d(x,y) < \infty$ for all $x,y \in G$) and locally finite (i.e. $|\{y \in G : \{x,y\} \in E\}| < \infty$ for all $x \in G$). For $x \in G$ and $r \geq 0$, denote $B(x,r) = \{y \in G : d(x,y) \leq r\}$.

Now assume that the graph G is endowed with a weight (conductance) μ_{xy}, which is a symmetric nonnegative function on $G \times G$ such that $\mu_{xy} > 0$ if and only if $x \sim y$. We call the pair (G,μ) a weighted graph. We can regard it as an electrical network. We define a quadratic form on (G,μ) as follows. Set

$$\mathcal{E}(f,g) = \frac{1}{2} \sum_{\substack{x,y \in G \\ x \sim y}} (f(x) - f(y))(g(x) - g(y))\mu_{xy} \qquad \text{for all} \quad f,g \in \mathbb{R}^G.$$

For each $x \in G$, let $\mu_x = \sum_{y \in G} \mu_{xy}$ and for each $A \subset G$, set $\mu(A) = \sum_{x \in A} \mu_x$. μ is a measure on G. Let $\{Y_n\}_{n \geq 0}$ be the discrete time Markov chain whose transition probabilities are given by

$$P(Y_{n+1} = y | Y_n = x) = \frac{\mu_{xy}}{\mu_x} =: P(x,y) \qquad \text{for all} \quad x,y \in G.$$

Y is is called a simple random walk when $\mu_{xy} = 1$ whenever $x \sim y$. The heat kernel of $\{Y_n\}_{n \geq 0}$ can be written as

$$p_n(x,y) := P^x(Y_n = y)/\mu_y \qquad \text{for all} \quad x,y \in G, \tag{3.3}$$

where we set $P^x(\cdot) := P(\cdot | Y_0 = x)$. Clearly, $p_n(x,y) = p_n(y,x)$. We sometimes consider a continuous time Markov chain $\{Y_t\}_{t \geq 0}$ with respect to μ which is defined as follows: each particle stays at a point, say x for (independent) exponential time with parameter 1, and then jumps to another point, say y with probability $P(x,y)$. The heat kernel for the continuous time Markov chain can be expressed as follows.

$$p_t(x,y) = P^x(Y_t = y)/\mu_y = \sum_{n=0}^{\infty} e^{-t} \frac{t^n}{n!} p_n(x,y) \qquad \text{for all} \quad x,y \in G.$$

The discrete Laplacian corresponding to $\{Y_t\}_{t \geq 0}$ is

$$\mathcal{L}f(x) = \sum_{\substack{y \in G \\ y \sim x}} P(x,y)f(y) - f(x) = \frac{1}{\mu_x} \sum_{\substack{y \in G \\ y \sim x}} \Big(f(y) - f(x)\Big)\mu_{xy}.$$

In this section, we assume the following condition on the weighted graph.

Definition 3.1. Let (G,μ) be a weighted graph.

(i) We say (G,μ) has controlled weights if there exists $p_0 > 0$ such that

$$P(x,y) = \mu_{xy}/\mu_x \geq p_0 \qquad \text{for all} \quad x \sim y \in G.$$

(ii) We say (G,μ) satisfies a volume doubling condition (VD) if there exists $c_1 > 1$ such that

$$\mu(B(x,2R)) \leq c_1 \mu(B(x,R)) \qquad \text{for all} \quad x \in G, R \geq 1. \tag{3.4}$$

3.2. Stability. We first introduce two types of perturbations.

Definition 3.2. Let $(G_1, \mu_1), (G_2, \mu_2)$ be weighted graphs with controlled weights.

(i) We say (G_2, μ_2) is a bounded perturbation of (G_1, μ_1) if $G_1 = G_2$ and there exist $c_1, c_2 > 0$ such that $c_1(\mu_1)_{xy} \le (\mu_2)_{xy} \le c_2(\mu_1)_{xy}$ for all $x \sim y$.

(ii) A map $T : G_1 \to G_2$ is called a rough isometry if there exist positive constants $c_1, \cdots, c_4 > 0$ such that the following holds for all $x, y \in G_1$ and $y' \in G_2$.

$$c_1^{-1} d_1(x,y) - c_2 \le d_2(T(x), T(y)) \le c_1 d_1(x,y) + c_2$$
$$d_2(T(G_1), y') \le c_3, \quad c_4^{-1}(\mu_1)_x \le (\mu_2)_{T(x)} \le c_4(\mu_1)_x.$$

where $d_i(\cdot,\cdot)$ is the the graph distance of (G_i, μ_i), for $i = 1, 2$. $(G_1, \mu_1), (G_2, \mu_2)$ are said to be rough isometric if there is a rough isometry between them.

The notion of rough isometry was first introduced by Kanai (1985). Note that rough isometry corresponds to (coarse) quasi-isometry in the field of geometric group theory, which was introduced by Gromov (1981).

We now define some (functional) inequalities.

Definition 3.3. Let (G, μ) be a weighted graph with controlled weights and let $\beta > 1$.

(i) We say (G, μ) satisfies sub-Gaussian heat kernel estimates (HK(β)) if there exist $c_1, \cdots, c_4 > 0$ such that for $x, y \in G, n \ge d(x,y) \vee 1$, the following holds:

$$p_n(x,y) \le \frac{c_1}{\mu(B(x, n^{1/\beta}))} \exp\Big(- c_2 \Big(\frac{d(x,y)^\beta}{n} \Big)^{1/(\beta-1)} \Big),$$
$$p_n(x,y) + p_{n+1}(x,y) \ge \frac{c_3}{\mu(B(x, n^{1/\beta}))} \exp\Big(- c_4 \Big(\frac{d(x,y)^\beta}{n} \Big)^{1/(\beta-1)} \Big).$$

(ii) We say (G, μ) satisfies (PI(β)), a scaled Poincaré inequality with exponent β, if there exists a constant $c_1 > 0$ such that for any ball $B_R := B(x_0, R) \subset G$ with $x_0 \in G$, $R \ge 1$ and $f : B_R \to \mathbb{R}$,

$$\sum_{x \in B_R} (f(x) - \bar{f}_{B_R})^2 \mu_x \le c_1 R^\beta \sum_{x \in B_R} \Gamma(f,f)(x).$$

Here $\bar{f}_{B_R} := \mu(B_R)^{-1} \sum_{y \in B_R} f(y)\mu_y$, and $\Gamma(f,f)(x) := \sum_{y \sim x} (f(x) - f(y))^2 \mu_{xy}$.

(iii) We say (G, μ) satisfies (CSA(β)), a cut-off Sobolev inequality in annuli with exponent β, if there exist a constant $c_1 > 0$ such that for every $x_0 \in G, R, r \ge 1$, there exists a cut-off function φ satisfying the following properties:

(a) $\varphi(x) = 1$ if $x \in B_R$, $\varphi(x) = 0$ if $x \in B_{R+r}^c$.

(b) Let $U = B_{R+r} \setminus B_R$. For any $f : U \to \mathbb{R}$,

$$\sum_{x \in U} f(x)^2 \Gamma(\varphi, \varphi)(x) \le c_1 \Big(\sum_{x \in U} \varphi(x)^2 \Gamma(f,f)(x) + r^{-\beta} \sum_{x \in U} f(x)^2 \mu_x \Big).$$

Theorem 3.4 ([2, 8, 9]). *Let (G, μ) be a weighted graph with controlled weights. Then,*

$$(\mathrm{VD}) + (\mathrm{PI}(\beta)) + (\mathrm{CSA}(\beta)) \Leftrightarrow (\mathrm{PHI}(\beta)) \Leftrightarrow (\mathrm{HK}(\beta)). \tag{3.5}$$

Here and in the following, (PHI(β)) means the discrete version of (PHI(d_w)) in Theorem 2.2 with $d_w = \beta$.

Remark 3.5.

(i) There are various other equivalent conditions to (3.5); see [26, 45] and references therein.

(ii) When one of (thus all) the above conditions holds, then it turns out that $\beta \geq 2$.

(iii) (CSA(2)) always holds in the graph context. (Take $\varphi(x) = 1 \wedge r^{-1} d(x, B^c_{R+r})$.) Thus Theorem 3.4 is an extension of (3.2) to the cases of $\beta > 2$ for graphs.

(iv) The main theorem in [2] is the equivalence of the upper bound of (HK(β)) and (CSA(β)) plus the Faber-Krahn inequality with exponent β. The results are stated on metric measure spaces.

For the $\beta = 2$ case, there is a well-known method called Moser's iteration to deduce the Harnack inequality in (3.2). In order for the method to work, it is necessary that the correct order can be deduced using linear cut-off functions. If we adopt similar arguments using the Lipschitz cut-off functions for the $\beta > 2$ case, then the estimates obtained are not sharp enough to establish the Harnack inequality. Roughly speaking, (CSA(β)) guarantees the existence of nice cut-off functions φ that satisfy $\mathcal{E}(\varphi, \varphi) \leq c_1 R^{-\beta} \mu(B_R)$. (Note that the order of the energy for the Lipschitz continuous cut-off function is $R^{-2}\mu(B_R)$.) The idea of the proof of the Harnack inequality when $\beta > 2$ is to apply Moser's iteration for weighted measures $\nu_x := \mu_x + R^\beta \Gamma(\varphi, \varphi)(x)$ using (CSA(β)).

Clearly, (VD), (PI(β)) and (CSA(β)) are stable under bounded perturbations. Further, it can be proved that they are stable under rough isometry (Hambly-Kumagai (2004)). We thus obtain the stability of (PHI(β)) and (HK(β)).

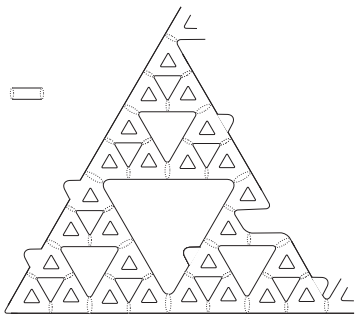


Figure 3.1. Fractal-like manifold

As mentioned above, Theorem 3.4 holds in the framework of metric measure spaces with local regular Dirichlet forms (especially Riemannian manifolds). It also holds when the walk dimension β is different for short times and long times. Figure 3.1 is a 2-dimensional Riemannian manifold whose global structure is like that of the gasket. This can be constructed from the left of Figure 2.1 by changing each bond to a cylinder and putting projections and

dents locally. The diffusion corresponding to the Dirichlet form moves on the surfaces of the cylinders. Using the generalization of Theorem 3.4, one can show that any divergence operator $\mathcal{L} = \sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}(a_{ij}(x)\frac{\partial}{\partial x_j})$ on the manifold which satisfies the uniform elliptic condition obeys (PHI(2)) for $R \le 1$ and (PHI($\log 5/\log 2$)) for $R \ge 1$.

3.3. Strongly recurrent case. The problem with Theorem 3.4 is that it is in general very difficult to check (CSA(β)). Under a stronger volume growth condition, a simpler equivalent condition is known.

For each $x \neq y \in G$, define the effective resistance between them by

$$R_{\mathrm{eff}}(x,y)^{-1} = \inf\left\{\mathcal{E}(f,f) : f(x) = 1, f(y) = 0, f \in \mathbb{R}^G\right\}. \tag{3.6}$$

We define $R_{\mathrm{eff}}(x,x) = 0$ for $x \in G$.

Definition 3.6.

(i) We say (G,μ) satisfies the volume growth condition (VG(β_-)) if there exist $K > 1$, $c_1 > 0$ with $\log c_1/\log K < \beta$ such that

$$\mu(B(x,KR)) \le c_1\mu(B(x,R)) \quad \text{for all} \quad x \in G, R \ge 1.$$

(ii) We say (G,μ) satisfies (RE(β)), the effective resistance bounds with exponent β, if there exist $c_1, c_2 > 0$ such that

$$\frac{c_1 d(x,y)^\beta}{\mu(B(x,d(x,y)))} \le R_{\mathrm{eff}}(x,y) \le \frac{c_2 d(x,y)^\beta}{\mu(B(x,d(x,y)))} \quad \text{for all} \quad x,y \in G.$$

Theorem 3.7. ([10]) *Let (G,μ) be a weighted graph with controlled weights and assume* (VG(β_-)). *Then,*

$$(\mathrm{RE}(\beta)) \Leftrightarrow (\mathrm{PHI}(\beta)) \Leftrightarrow (\mathrm{HK}(\beta)).$$

Under the above conditions, the Markov chain is strongly recurrent in the sense that there exists $p_1 > 0$ such that $P^x(\sigma_{\{y\}} < \sigma_{B(x,2r)^c}) \ge p_1$ for all $x \in G$, $r \ge 1$ and $y \in B(x,r)$, where $\sigma_A = \min\{n \ge 0 : Y_n \in A\}$. Theorem 3.7 is also generalized to the framework of metric measure spaces (Kigami ([34]), Kumagai (2004)).

One can refine the proof of this theorem to a statement which is applicable for random media as we discuss in the next section.

Open problem II. Provide a simpler equivalent condition to (HK(β)) that is applicable to a general graph.

4. Random walk on percolation clusters

From now on, we will discuss random walk on random media. We will consider a random weighted graph $(\mathcal{G}(\omega),\mu(\omega))$ for $\omega \in \Omega$. $(\Omega,\mathcal{F},\mathbb{P})$ is a probability space that governs randomness of the weighted graph. Note that we no longer have controlled weights and we cannot expect (VD) in general, so the arguments given in previous sections are not applicable directly. We are interested in the long time behavior of the corresponding Markov chain $\{Y_t^\omega\}_{t\ge 0}$ at the quenched level (i.e. $\mathbb{P}$-a.s. level); we are especially interested in the following two questions:

(Q1) Long time heat kernel estimates for $p_t^\omega(\cdot,\cdot)$.

(Q2) Scaling limit of $\{Y_t^\omega\}_{t\geq 0}$.

(Recall that the suffix ω stands for the randomness of the media.) The prototypical example is random walk on percolation clusters on $\mathbb{Z}^d$, $d \geq 2$.

4.1. Supercritical case. We first consider the supercritical case. In this case, $\{\mu_e : e \in E_d\}$ are Bernoulli random variables; $\mathbb{P}(\mu_e = 1) = p, \mathbb{P}(\mu_e = 0) = 1-p$ where $p > p_c(\mathbb{Z}^d)$ – see Section 1 for the definition of $p_c(\mathbb{Z}^d)$. We know that there exists a unique infinite connected component of edges with conductance 1, which we denote by $\mathcal{G}(\omega)$. We will condition on the event $\{0 \in \mathcal{G}(\omega)\}$ and define $\mathbb{P}_0(\cdot) := \mathbb{P}(\cdot|0 \in \mathcal{G})$.

As for (Q1), the following heat kernel estimates are proved in [6].

Theorem 4.1. *There exist constants $\eta, c_1, \cdots, c_6 > 0$ and a family of random variables $\{U_x\}_{x\in\mathbb{Z}^d}$ with $\mathbb{P}(U_x \geq n) \leq c_1 \exp(-c_2 n^\eta)$ such that the following holds $\mathbb{P}_0$-a.s. for $t \geq U_x \vee |x-y|$:*

$$c_3 t^{-d/2} \exp(-c_4|x-y|^2/t) \leq p_t^\omega(x,y) \leq c_5 t^{-d/2} \exp(-c_6|x-y|^2/t). \tag{4.1}$$

The proof uses (3.2) in spirit. A ball $B(x,r)$ is said to be "good" if the volume is comparable to r^d and (PI(2)) holds for the ball. It is proved that a ball is good with high probability and the Borel-Cantelli lemma is used to establish some quenched estimates. Part of the proof of (3.2) is used to establish some heat kernel estimates on good balls.

As for (Q2), it turns out that the quenched invariance principle holds, namely $\varepsilon Y^\omega_{t/\varepsilon^2}$ converges as $\varepsilon \to 0$ to Brownian motion on $\mathbb{R}^d$ (with covariance $\sigma^2 I$, $\sigma > 0$) $\mathbb{P}_0$-a.e. ω. This was first proved in [43] for $d \geq 4$ and later extended to all $d \geq 2$ in [18, 40]. The proof for $d \geq 3$ uses Theorem 4.1.

Theorem 4.2. *$\mathbb{P}_0$-a.s., $\varepsilon Y_{t/\varepsilon^2}$ converges (under P^0_ω) in law to Brownian motion on $\mathbb{R}^d$ with covariance $\sigma^2 I$ where $\sigma > 0$ is a non-random constant.*

Furthermore, the quenched local limit theorem holds for this model ([12]).

Let us emphasize that percolation provides one of the natural degenerate models in the sense that uniform ellipticity does not hold, and it is a highly non-trivial fact that the scaling limit is Brownian motion with probability one. For the random conductance model discussed below, when $\mathbb{E}\mu_e < \infty$, a weak form of convergence was already proved in the 1980s that the convergence holds in law under $\mathbb{P}_0 \times P^0_\omega$; a milestone by Kipnis-Varadhan (1986). (See also De Masi-Ferrari-Goldstein-Wick (1989) and Kozlov (1985).) This is sometimes referred to as the annealed (or averaged) invariance principle. It took about two decades to improve the annealed invariance principle to the quenched one.

Remark 4.3. More generally, (Q1) and (Q2) have been extensively studied on the random conductance model. Let $\{\mu_e : e \in E_d\}$ be stationary ergodic that takes non-negative values, and assume $\mathbb{P}(\mu_e > 0) > p_c(\mathbb{Z}^d)$. Then there exists a unique infinite connected component of edges with positive conductance, which we denote by $\mathcal{G}(\omega)$. The random weighted graph $(\mathcal{G}, \mu)$ is the random conductance model. For the i.i.d. case, although there are examples where the heat kernel behaves anomalously (Berger-Biskup-Hoffman-Kozma (2008)), it is proved that quenched invariance principle as in Theorem 4.2 holds; further, $\sigma > 0$ is non-random if $\mathbb{E}\mu_e < \infty$ and $\sigma = 0$ (i.e. the limiting process does not

move) if $\mathbb{E}\mu_e = \infty$ (Biskup-Prescott (2007), Mathieu-Piatnitski (2007), Barlow-Deuschel (2010), Andres-Barlow-Deuschel-Hambly (2013)). When $\mathbb{P}(\mu_e \geq u) \sim u^{-\alpha}$as $u \to \infty$ for $\alpha \in (0,1)$, a special case of $\mathbb{E}\mu_e = \infty$, a suitably rescaled Markov chain converges to an anomalous process. It converges to the Fractional-Kinetics (FK) process when $d \geq 2$, where the corresponding heat kernel obeys a fractional time heat equation, and to the Fontes-Isopi-Newman (FIN) diffusion when $d = 1$ (Barlow-Černý (2011), Černý (2011)). See [19, 36] for details. For general ergodic media with $\mathbb{P}(0 < \mu_e < \infty) = 1$, Andres-Deuschel-Slowik ([3]) has proved the quenched invariance principle under some integrability condition of the media. They use Moser's iteration instead of the heat kernel estimates. See Procaccia-Rosenthal-Sapozhnikov (2013) for the quenched invariance principle on a class of degenerate ergodic media such as random interlacements.

4.2. Critical case. We next consider random walk on percolation clusters at criticality. If $d = 2$ or $d \geq 19$ (or $d > 6$ for spread-out models mentioned below) it is known that $\theta(p_c) = 0$, i.e. there is no infinite open cluster $\mathbb{P}$-a.s.; see for example [27]. (Fitzner-van der Hofstad (2014) extends $d \geq 19$ to $d \geq 15$.) It is conjectured that this holds for $d \geq 2$. However, when $p = p_c$, in any box of side n there exist with high probability open clusters of diameter of order n. In order to study mesoscopic properties of these large finite clusters, we will regard them as subsets of an infinite cluster $\mathcal{G}$, called the incipient infinite cluster (IIC for short) and analyze the IIC. This IIC $\mathcal{G} = \mathcal{G}(\omega)$ is our random graph.

The IIC was constructed when $d = 2$ in [31], by taking the limit as $N \to \infty$ of the cluster $\mathcal{C}(0)$ conditioned to intersect the boundary of a box of side N centered at the origin. For large d, a construction of the IIC in $\mathbb{Z}^d$ is given in van der Hofstad-Járai (2004), using the lace expansion. (The results are believed to hold for any $d > 6$.) They also prove the existence and some properties of the IIC for all $d > 6$ for spread-out models: these include the case when there is a bond between x and y with probability pL^{-d} whenever y is in a cube side L with center x, and the parameter L is large enough. The IIC measure can be written as follows:

$$\mathbb{P}_{\mathrm{IIC}}(F) = \lim_{d(0,x)\to\infty} \mathbb{P}_{p_c}(F | 0 \leftrightarrow x) \qquad \text{for all} \quad F : \text{cylindrical event}, \tag{4.2}$$

where $\{0 \leftrightarrow x\}$ is the event that 0 and x are in the same open cluster. In the following, we will write $\mathcal{G} = \mathcal{G}_d(\omega)$ for the IIC in $\mathbb{Z}^d$. It is believed that the global properties of $\mathcal{G}$ are the same for all $d > d_c$, both for nearest neighbor and spread-out models, where d_c is the *critical dimension* which is 6 for the percolation model.

Let $Y = \{Y_n^\omega\}_{n\in\mathbb{N}}$ be simple random walk on $\mathcal{G}$, and $p_n^\omega(x,y)$ be its heat kernel. The Alexander-Orbach conjecture mentioned in the introduction can be stated as follows: for any $d \geq 2$, $d_s(\mathcal{G}) = 4/3$, $\mathbb{P}_{\mathrm{IIC}}$-a.e., where d_s was defined in (1.1).

The Alexander-Orbach conjecture turns out to be true on a high dimensional percolation cluster ([35]) as we state in the following.

Theorem 4.4. *There exists $\alpha > 0$ such that the following holds when $d > 6$ for the spread-out model ($d \geq 19$ for the nearest neighbor model): For $\mathbb{P}_{\mathrm{IIC}}$-a.e. $\omega \in \Omega$ and $x \in \mathcal{G}(\omega)$, there exist $N_x(\omega), R_x(\omega) \in \mathbb{N}$ such that*

$$(\log n)^{-\alpha} n^{-\frac{2}{3}} \leq p_{2n}^\omega(x,x) \leq (\log n)^{\alpha} n^{-\frac{2}{3}} \qquad \textit{for all} \quad n \geq N_x(\omega), \tag{4.3}$$

$$(\log R)^{-\alpha} R^3 \leq E_\omega^x \tau_{B(0,R)} \leq (\log R)^{\alpha} R^3 \qquad \textit{for all} \quad R \geq R_x(\omega), \tag{4.4}$$

where $\tau_A := \min\{n \geq 0 : Y_n \notin A\}$.

In the next subsection, we will briefly discuss how this was proved.

4.2.1. Heat kernel estimates on random media. As we mentioned in the end of the last section, Theorem 3.7 (especially its proof) turns out to be useful even for random walk on random media. Below we give a general theorem.

Let $(\mathcal{G}(\omega), \omega \in \Omega)$ be a random graph on $(\Omega, \mathcal{F}, \mathbb{P})$; for $\mathbb{P}$-a.e. ω, we assume that $\mathcal{G}(\omega)$ is a connected locally finite graph that contains a distinguished point $0 \in \mathcal{G}(\omega)$. For each ω, we put conductance 1 for each bond and let $\{Y_n^\omega\}$ be the simple random walk on $\mathcal{G}$. Let $B(0,R)$ be the ball of radius R centered at 0 with respect to the graph distance $d(\cdot,\cdot)$. For $D, \lambda \geq 1$, we say $B(0,R)$ in $\mathcal{G}$ is λ-good if

$$\frac{R^D}{\lambda} \leq \mu(B(0,R)) \leq \lambda R^D, \quad \frac{R}{\lambda} \leq R_{\mathrm{eff}}(0, B(0,R)^c). \tag{4.5}$$

Here $R_{\mathrm{eff}}(\cdot,\cdot)$ is the effective resistance defined in (3.6). The following are the general estimates in [13, 37].

Theorem 4.5. *If there exist $R_0, \lambda_0 \geq 1$ and $q_0 > 0$ such that*

$$\mathbb{P}(\{\omega : B(0,R) \text{ is } \lambda\text{-good } \}) \geq 1 - \lambda^{-q_0} \qquad \text{for all} \quad R \geq R_0, \lambda \geq \lambda_0, \tag{4.6}$$

then there exists $c > 0$ such that the following holds:
For $\mathbb{P}$-a.e. $\omega \in \Omega$ and $x \in \mathcal{G}(\omega)$, there exist $N_x(\omega), R_x(\omega) \in \mathbb{N}$ such that

$$(\log n)^{-c} n^{-\frac{D}{D+1}} \leq p^\omega_{2n}(x,x) \leq (\log n)^c n^{-\frac{D}{D+1}} \qquad \text{for all } \ n \geq N_x(\omega), \tag{4.7}$$

$$(\log R)^{-c} R^{D+1} \leq E^x_\omega \tau_{B(0,R)} \leq (\log R)^c R^{D+1} \qquad \text{for all } \ R \geq R_x(\omega). \tag{4.8}$$

In particular, $d_s(\mathcal{G}(\omega)) = \frac{2D}{D+1}$, $\mathbb{P}$–a.s. ω, and the random walk is recurrent.

Furthermore, if (4.6) *holds with $\exp(-c_1 \lambda^{q_0})$ instead of λ^{-q_0}, then* (4.7) *and* (4.8) *hold with $(\log\log\,\cdot)^{\pm c}$ instead of $(\log\,\cdot)^{\pm c}$.*

In the above statement, the volume growth is of order R^D and the resistance growth is linear. In [37], a general version is given where both growths are controlled by increasing functions with $c_1(R/r)^{\beta_1} \leq f(R)/f(r) \leq c_2(R/r)^{\beta_2}$ for $0 < r < R$, where $0 < \beta_1 \leq \beta_2$ are constants. For this general version, we need to add an extra condition $R_{\mathrm{eff}}(0,z) \leq \lambda f(d(0,z))$ for all $z \in B(0,R)$ in (4.5). Note that this extra condition is always true for the linear case.

Open problem III. Provide a simpler sufficient condition for the heat kernel and exit time estimates for $d_s \geq 2$.

4.2.2. Applying Theorem 4.5 to concrete models. In [35], the condition (4.6) is proved using the control of the two-point function that can be obtained using the lace expansion. Write $x \leftrightarrow y$ if x and y are connected by open edges.

Proposition 4.6. *For the critical bond percolation, assume that the following holds:*

$$c_1|x|^{2-d} \leq \mathbb{P}_{p_c}(0 \leftrightarrow x) \leq c_2|x|^{2-d} \qquad \text{for all} \quad x \in \mathcal{G}(\omega). \tag{4.9}$$

Then (4.6) *in Theorem 4.5 holds for $\mathbb{P}_{\mathrm{IIC}}$ with $D = 2$.*

When d is high enough, (4.9) is proved using the lace expansion (Hara-van der Hofstad-Slade (2003) for $d > 6$ for the spread-out model, Hara (2008) for $d \geq 19$ for the nearest neighbor model), which implies Theorem 4.4.

There are other models where anomalous behavior of random walk has been proved by verifying (4.6) in Theorem 4.5. We list up some of them. For (i)-(iii), $D = 2$ and $d_s = 4/3$. For (i), (4.6) holds with $\exp(-c_1\lambda^{q_0})$ instead of λ^{-q_0}.

(i) IIC for critical percolation on regular trees ([14]).

(ii) IIC for spread out oriented percolation for $d \geq 6$ ([13]).

(iii) Invasion percolation on a regular tree ([4]).

(iv) IIC for α-stable Galton-Watson trees conditioned to survive forever (Croydon-Kumagai (2008)): $D = \alpha/(\alpha - 1)$ and $d_s = 2\alpha/(2\alpha - 1)$.

(v) 2-dimensional uniform spanning trees ([15]): $D = 8/5$ and $d_s = 16/13$ – See Section 5.2 for details.

[28] partly generalized the results in [35], and proved the Alexander-Orbach conjecture for the IIC in high dimensions, both for long-range and finite-range percolation.

For the model (i), we have much more detailed estimates ([14]).

Theorem 4.7. *The heat kernel of simple random walk on the IIC for critical percolation on the regular tree obeys the following estimates.*

(i) *(4.3) and (4.4) hold with* $(\log\log\,\cdot)^{\pm\alpha}$ *instead of* $(\log\,\cdot)^{\pm\alpha}$.

(ii) *It holds that for* $\mathbb{P}_{\mathrm{IIC}}$*-a.e.* ω

$$\liminf_{n\to\infty}(\log\log n)^{1/6}n^{2/3}p^{\omega}_{2n}(0,0) \leq 2.$$

(iii) *The annealed heat kernel* $\mathbb{E}_{\mathrm{IIC}}[p_{2n}(x,y)|x,y \in \mathcal{G}]$ *obeys the sub-Gaussian estimates* (2.1) *with* $d_f = 2, d_w = 3$ *for* $n \geq d(x,y) \vee 1$.

As we have seen above, the quenched estimates have oscillation of $\log\log$ order whereas the annealed estimates do not. Detailed off-diagonal heat kernel estimates (which hold with high probability) are also obtained in [14, Theorem 4.9, 4.10].

4.2.3. Below critical dimensions. For low dimensions, there are only a few rigorous results. One of the most attractive models is the IIC for 2-dimensional critical percolation. In [32], Kesten proves sub-diffusive behavior of simple random walk on the IIC for 2-dimensional critical percolation cluster (also shows the existence of the IIC in [31]). Namely, let $\{Y^{\omega}_n\}_{n\geq 0}$ be a simple random walk on the IIC, then there exists $\epsilon > 0$ such that the $\mathbb{P}_{\mathrm{IIC}}$-distribution of $n^{-\frac{1}{2}+\epsilon}d(0,Y_n)$ is tight. A quenched version of Kesten's result is established both for the IIC and the invasion percolation cluster (Damron-Hanson-Sosoe (2013)). For bond percolation on $\mathbb{Z}^d$, the critical dimension is 6. The Alexander-Orbach conjecture is considered to be false for $d \leq 5$ and some numerical simulations (cf. [17], [29, Section 7.4]) support this. It is a challenging problem to prove this rigorously, especially for $d = 2$.

It is proved in [30] that the effective resistance between the origin and generation n of the incipient infinite oriented branching random walk in $d < 6$ is $O(n^{1-\gamma})$ for some $\gamma > 0$. It is interesting to see that, while the critical dimension of the model is 4, asymptotic behavior

of the random walk changes already at $d = 5$. The precise resistance exponent (even its existence) is not known.

Other low dimensional random media for which heat kernel/exit time estimates have been studied include the uniform infinite planar triangulation (Benjamini-Curien (2013); see also Gurel-Gurevich and Nachmias (2013)), the critical percolation cluster for the diamond lattice (Hambly-Kumagai (2010)), and the non-intersecting two-sided random walk trace on $\mathbb{Z}^2$ and $\mathbb{Z}^3$ (Shiraishi (2014+)). See [36, Section 7.4] for details.

Open problem IV. (i) Prove the existence of d_s and d_w for lower dimensional models. Disprove (or prove) the Alexander-Orbach conjecture for the models.
(ii) Compute resistance for random media when the resistance growth is not linear.

Remark 4.8. Heat kernel estimates and scaling limits have been considered for random walks on the long-range percolation model and its variants. See [20, 21] and references therein.

5. Scaling limits of random walks on random media

In this section, we discuss (Q2) (i.e. question about scaling limits of random walks) for random media. It is proved by Croydon (2008) that the distribution of the rescaled simple random walk on critical finite variance Galton-Watson tree converges to Brownian motion on the Aldous tree (see Croydon (2010) for the infinite variance case). Below, we give two more examples.

5.1. Erdős-Rényi random graph in critical window. Let $V_N := \{1, 2, \cdots, N\}$. The Erdős-Rényi random graph is a percolation on the complete graph with vertices in V_N, namely each bond $\{i, j\}$, $i, j \in V_N$ is open with probability $p \in [0, 1]$ and closed otherwise, independently of all the others. Denote its largest connected component by $\mathcal{C}^N$. It is known that this model exhibits a phase transition around $p \sim c/N$ in that the following holds with high probability (Erdős-Rényi (1960)):

$$c < 1 \Rightarrow |\mathcal{C}^N| = O(\log N),\ c > 1 \Rightarrow |\mathcal{C}^N| \asymp N,\ c = 1 \Rightarrow |\mathcal{C}^N| \asymp N^{2/3}.$$

We will consider finer scaling (the so-called critical window), namely we will take $p = 1/N + \lambda N^{-4/3}$ for fixed $\lambda \in \mathbb{R}$. In this window, the size of the i-th largest connected component is of order $N^{2/3}$ for each $i \in \mathbb{N}$. The following results hold for each i-th largest connected component; for simplicity, we state them for the $\mathcal{C}^N$.

There exists a random compact metric space $\mathcal{M} = \mathcal{M}_\lambda$ such that the following holds in the Gromov-Hausdorff sense

$$N^{-1/3}\mathcal{C}^N \xrightarrow{d} \mathcal{M},$$

where $\mathcal{C}^N$ is considered as a rooted metric space (Addario-Berry, Broutin and Goldschmidt (2012); see also Aldous (1997)). The concrete construction of $\mathcal{M}$ is also known. Let $\{Y_m^{\mathcal{C}^N}\}_{m \geq 0}$ be the simple random walk on $\mathcal{C}^N$. Then the following holds.

Theorem 5.1 ([22]).

(i) *There exist Brownian motion* $\{B_t^{\mathcal{M}}\}_{t\geq 0}$ *on* $\mathcal{M}$ *such that*

$$\{N^{-1/3}Y^{\mathcal{C}^N}_{[Nt]}\}_{t\geq 0} \xrightarrow{d} \{B_t^{\mathcal{M}}\}_{t\geq 0}, \quad \mathbb{P}-a.s.$$

(ii) *There exist a jointly continuous heat kernel* $p_t^{\mathcal{M}}(\cdot,\cdot)$ *of Brownian motion and* $\theta, T_0, c_1, \cdots, c_4 > 0$ *such that for* $\mathbb{P}$*-a.e.* $\omega \in \Omega$*,*

$$p_t^{\mathcal{M}}(x,y) \leq c_1 t^{-\frac{d_f}{d_w}} \ell(t^{-1})^{\theta} \exp\left\{-c_2\left(\frac{d(x,y)^{d_w}}{t}\right)^{\frac{1}{d_w-1}} \ell\left(\frac{d(x,y)}{t}\right)^{-\theta}\right\} \tag{5.1}$$

$$p_t^{\mathcal{M}}(x,y) \geq c_3 t^{-\frac{d_f}{d_w}} \ell(t^{-1})^{-\theta} \exp\left\{-c_4\left(\frac{d(x,y)^{d_w}}{t}\right)^{\frac{1}{d_w-1}} \ell\left(\frac{d(x,y)}{t}\right)^{\theta}\right\} \tag{5.2}$$

for all $x, y \in \mathcal{M}, t \leq T_0$ *with* $\ell(x) := 1 \vee \log x$ *and* $d_f = 2, d_w = 3$*.*

It is known that the L^p-mixing time of the simple random walk on $\mathcal{C}^N$ converges in $\mathbb{P}$-distribution to that of Brownian motion on $\mathcal{M}$ (Croydon-Hambly-Kumagai (2012); see also Nachmias-Peres (2008)).

5.2. 2-dimensional uniform spanning tree. Let $\Lambda_n := [-n,n]^2 \cap \mathbb{Z}^2$, which we consider as a graph with edges between lattice neighbors. A spanning tree of Λ_n is a subgraph that connects all the vertices of Λ_n and contains no cycles. Let $\mathcal{U}^{(n)}$ be a spanning tree of Λ_n selected uniformly at random from all possibilities. Pemantle (1991) showed that one could then define a uniform spanning tree (UST) of $\mathbb{Z}^2$, which we denote by $\mathcal{U}$, as the local limit of $\mathcal{U}^{(n)}$ as $n \to \infty$. He also showed that the distribution of $\mathcal{U}$ is independent of the boundary conditions (such as wired, free) on Λ_n. An alternative and very useful construction of $\mathcal{U}$ involves Wilson's algorithm (1996), which can be described as follows. Enumerate $\mathbb{Z}^2$ arbitrarily as $x_0, x_1, \cdots$ and let $\mathcal{U}(0) = \{x_0\}$. For $k \geq 1$, given $\mathcal{U}(k-1)$, run the loop-erased random walk (LERW) from x_k to $\mathcal{U}(k-1)$ and define $\mathcal{U}(k)$ to be the union of the path and $\mathcal{U}(k-1)$. (Here, LERW is a process introduced by Lawler (1980) which is obtained by chronologically erasing loops from the simple random walk.) We then obtain $\mathcal{U} = \cup_{k\geq 0}\mathcal{U}(k)$ – see [39] for more details about the UST.

Now, let M_n be the number of steps of the loop-erasure of a simple random walk on $\mathbb{Z}^2$ from 0 to the circle of radius n. It follows from Lawler (2013) that $E^0 M_n \asymp n^{5/4}$ (note that $\lim_{n\to\infty} \log E^0 M_n / \log n = 5/4$ was shown by Kenyon (2000)). Applying this in conjunction with Wilson's algorithm, it has been established that $|B_{\mathcal{U}}(0,R)| \asymp R^{2/(5/4)} = R^{8/5}$ with high probability where $B_{\mathcal{U}}(x,R)$ is the ball with respect to the graph distance. In particular, in [15], the condition of Theorem 4.5 is proved with $D = 8/5$, as mentioned in Section 4.2.2.

In the seminal paper by Schramm (2000), the topological properties of any possible scaling limit of the 2-dimensional UST $\mathcal{U}$ were investigated. (The uniqueness of the scaling limit for a UST in a 2-dimensional domain was established in Lawler-Schramm-Werner (2004).) In [11], the convergence of $\mathcal{U}$ is discussed in terms of the generalized Gromov-Hausdorff-Prohorov topology. It is proved that the law of the UST is tight under rescaling in a space of measured, rooted real trees embedded into Euclidean space. Let $\mathcal{T}$ be the limiting real tree when the lattice spacing is rescaled using the subsequence $\{\delta_i\}_{i\geq 1}$, $\rho_{\mathcal{T}}$ be its root, $\phi_{\mathcal{T}}$ be the random embedding of $\mathcal{T}$ into $\mathbb{R}^2$, and $X^{\mathcal{T}}$ be Brownian motion on $\mathcal{T}$ started from

$\rho_{\mathcal{T}}$. Then the following holds, where we write $X^{\mathcal{U}}$ for the simple random walk on $\mathcal{U}$ started from 0.

Theorem 5.2 ([11]). *The annealed law of* $\{(\delta_i X^{\mathcal{U}}_{\delta_i^{-13/4}t} : t \geq 0)\}_{i\geq 1}$ *converges to the annealed law of* $\phi_{\mathcal{T}}(X^{\mathcal{T}})$. *Furthermore, there exists a jointly continuous heat kernel* $p^{\mathcal{T}}_t(\cdot,\cdot)$ *of* $X^{\mathcal{T}}$ *such that, for each* $R > 0$ *and* $\mathbb{P}$*-a.e.* $\omega \in \Omega$, *one can find* $T_0 > 0$ *such that* (5.1) *and* (5.2) *hold for all* $x, y \in B_{\mathcal{T}}(\rho_{\mathcal{T}}, R)$, $t \leq T_0$ *with* $\ell(x) := 1 \vee \log x$ *and* $d_f = 8/5, d_w = d_f + 1 = 13/5$.

Note that the exponent $13/4 = (5/4) \cdot d_w$ above is the walk dimension with respect to the Euclidean distance.

6. Conclusions

We have provided an overview of the stream of research on anomalous random walks and diffusions. Through the detailed study of diffusions on exactly self-similar fractals, it became apparent that Brownian motion on fractals typically obeys sub-Gaussian heat kernel estimates. This motivated the development of stability theory for such anomalous diffusions/random walks which is a generalization of the classical perturbation theory of Gaussian bounds. Then, some of the results in this direction turned out to be useful in analyzing random walks in random media. Although not discussed in this paper, such a stability theory also gives new insights to analysis on metric measure spaces.

There are many interesting random media whose dynamical properties are not yet known. Necessity is the Mother of Invention. We believe that further developments will continue to lead to important interactions between probability, analysis and mathematical physics.

Acknowledgements. The author thanks Martin Barlow, David Croydon, Naotaka Kajino and Gordon Slade for valuable comments on a draft of this paper. This research was partially supported by the Grant-in-Aid for Scientific Research (A) 25247007, Japan.

References

[1] Alexander, S. and Orbach, R., *Density of states on fractals: "fractons"*, J. Physique (Paris) Lett. **43** (1982), L625–L631.

[2] Andres, S. and Barlow, M.T., *Energy inequalities for cutoff functions and some applications*, J. reine angew. Math., to appear.

[3] Andres, S., Deuschel, J.-D., and Slowik, M., *Invariance principle for the random conductance model in a degenerate ergodic environment*, Ann. Probab., to appear.

[4] Angel, O., Goodman, J., den Hollander, F., and Slade, G., *Invasion percolation on regular trees*, Ann. Probab. **36** (2008), 420–466.

[5] Barlow, M.T., *Diffusions on fractals*, Lect. Notes in Math. **1690**, Springer, New York, 1998.

[6] ———, *Random walks on supercritical percolation clusters*, Ann. Probab. **32** (2004), 3024-3084.

[7] ———, *Analysis on the Sierpinski carpet*, Analysis and geometry of metric measure spaces, 27–53, CRM Proc. Lect. Notes **56**, Amer. Math. Soc., Providence, RI, 2013.

[8] Barlow, M.T. and Bass, R.F., *Stability of parabolic Harnack inequalities*, Trans. Amer. Math. Soc. **356** (2003), 1501–1533.

[9] Barlow M.T., Bass, R.F., and Kumagai, T., *Stability of parabolic Harnack inequalities on measure metric spaces*, J. Math. Soc. Japan **58** (2006), 485–519.

[10] Barlow, M.T., Coulhon, T., and Kumagai, T., *Characterization of sub-Gaussian heat kernel estimates on strongly recurrent graphs*, Comm. Pure Appl. Math. **58** (2005), 1642–1677.

[11] Barlow, M.T., Croydon, D., and Kumagai, T., *Subsequential scaling limits of simple random walk on the two-dimensional uniform spanning tree*, in preparation.

[12] Barlow, M.T. and Hambly, B.M., *Parabolic Harnack inequality and local limit theorem for random walks on percolation clusters*, Electron. J. Probab. **14** (2009), 1–27.

[13] Barlow, M.T., Járai, A.A., Kumagai, T., and Slade, G., *Random walk on the incipient infinite cluster for oriented percolation in high dimensions*, Comm. Math. Phys. **278** (2008), 385–431.

[14] Barlow, M.T. and Kumagai, T., *Random walk on the incipient infinite cluster on trees*, Illinois J. Math. **50** (2006), 33–65.

[15] Barlow, M.T. and Masson, R., *Spectral dimension and random walks on the two dimensional uniform spanning tree*, Comm. Math. Phys. **305** (2011), 23–57.

[16] Barlow, M.T. and Perkins, E.A., *Brownian motion on the Sierpiński gasket*, Probab. Theory Relat. Fields **79** (1988), 543–623.

[17] Ben-Avraham, D. and Havlin, S., *Diffusion and reactions in fractals and disordered systems*, Cambridge University Press, Cambridge, 2000.

[18] Berger, N. and Biskup, M., *Quenched invariance principle for simple random walk on percolation clusters*, *Probab. Theory Relat. Fields* **137** (2007), 83–120.

[19] Biskup, M., *Recent progress on the random conductance model*, Probability Surveys **8** (2011), 294–373.

[20] Chen, Z.-Q., Kim, P., and Kumagai, T., *Discrete approximation of symmetric jump processes on metric measure spaces*, Probab. Theory Relat. Fields **155** (2013), 703–749.

[21] Crawford, N. and Sly, A., *Simple random walks on long range percolation clusters I: heat kernel bounds*, Probab. Theory Relat. Fields **154** (2012), 753–786, *II: scaling limits*, Ann. Probab. **41** (2013), 445–502.

[22] Croydon, D.A., *Scaling limit for the random walk on the largest connected component of the critical random graph*, Publ. RIMS. Kyoto Univ. **48** (2012), 279–338.

[23] Fukushima, M., Oshima, Y., and Takeda, M., *Dirichlet forms and symmetric Markov processes*, de Gruyter, Berlin, 2011 (2nd Edition).

[24] De Gennes, P.G., *La percolation: un concept unificateur*, La Recherche **7** (1976), 919–927.

[25] Grigor'yan, A., *Heat kernel and analysis on manifolds*, Amer. Math. Soc., Providence, RI; International Press, Boston, MA, 2009.

[26] Grigor'yan, A. and Telcs, A., *Two-sided estimates of heat kernels on metric measure spaces*, Ann. Probab. **40** (2012), 1212–1284.

[27] Grimmett, G., *Percolation*, Springer, Berlin, 1999 (2nd Edition).

[28] Heydenreich, M., van der Hofstad, R., and Hulshof, T., *Random walk on the high-dimensional IIC*, Comm. Math. Phys. **329** (2014), 57-115.

[29] Hughes, B.D., *Random walks and random environments, volume 2: random environments*, Oxford University Press, Oxford, 1996.

[30] Járai, A.A. and Nachmias, A., *Electrical resistance of the low dimensional critical branching random walk*, arXiv:1305.1092.

[31] Kesten, H., *The incipient infinite cluster in two-dimensional percolation*, Probab. Theory Relat. Fields **73** (1986), 369–394.

[32] ———, *Subdiffusive behavior of random walk on a random cluster*, Ann. Inst. H. Poincaré Probab. Statist **22** (1986), 425–487.

[33] Kigami, J., *Analysis on fractals*, Cambridge Univ. Press, Cambridge, 2001.

[34] ———, *Resistance forms, quasisymmetric maps and heat kernel estimates*, Mem. Amer. Math. Soc. **216** (2012), no. 1015.

[35] Kozma, G. and Nachmias, A., *The Alexander-Orbach conjecture holds in high dimensions*, Invent. Math. **178** (2009), 635–654.

[36] Kumagai, T., *Random walks on disordered media and their scaling limits*, Lect. Notes in Math. **2101**, Springer, New York, 2014.

[37] Kumagai, T. and Misumi, J., *Heat kernel estimates for strongly recurrent random walk on random media*, J. Theoret. Probab. **21** (2008), 910–935.

[38] Kusuoka, S., *Diffusion processes on nested fractals*, Lect. Notes in Math. **1567**, Springer, New York, 1993.

[39] Lyons, R. and Peres, Y., *Probability on trees and networks*, Cambridge University Press, in preparation. Current version available at http://mypage.iu.edu/~rdlyons/.

[40] Mathieu, P. and Piatnitski, A., *Quenched invariance principles for random walks on percolation clusters*, Proc. Roy. Soc. A **463** (2007), 2287–2307.

[41] Rammal, R. and Toulouse, G., *Random walks on fractal structures and percolation clusters*, J. Physique Lettres **44** (1983), L13–L22.

[42] Saloff-Coste, L., *Aspects of Sobolev-type inequalities*, Cambridge Univ. Press, Cambridge, 2002.

[43] Sidoravicius, V. and Sznitman, A.-S., *Quenched invariance principles for walks on clusters of percolation or among random conductances*, Probab. Theory Relat. Fields **129** (2004), 219–244.

[44] Strichartz, R.S., *Differential equations on fractals: a tutorial*, Princeton University Press, Princeton, NJ, 2006.

[45] Telcs, A., *The art of random walks*, Lect. Notes in Math. **1885**, Springer, New York, 2006.

Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, Japan.
E-mail: kumagai@kurims.kyoto-u.ac.jp

The proximal distance algorithm

Kenneth Lange and Kevin L. Keys

Abstract. The MM principle is a device for creating optimization algorithms satisfying the ascent or descent property. The current survey emphasizes the role of the MM principle in nonlinear programming. For smooth functions, one can construct an adaptive interior point method based on scaled Bregmann barriers. This algorithm does not follow the central path. For convex programming subject to nonsmooth constraints, one can combine an exact penalty method with distance majorization to create versatile algorithms that are effective even in discrete optimization. These proximal distance algorithms are highly modular and reduce to set projections and proximal mappings, both very well-understood techniques in optimization. We illustrate the possibilities in linear programming, binary piecewise-linear programming, nonnegative quadratic programming, ℓ_0 regression, matrix completion, and inverse sparse covariance estimation.

Mathematics Subject Classification (2010). Primary 90C59; Secondary 65C60.

Keywords. Majorization, convexity, exact penalty method, computational statistics.

1. Introduction

The MM principle is a device for constructing optimization algorithms [4, 25, 28–30]. In essence, it replaces the objective function $f(\boldsymbol{x})$ by a simpler surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ anchored at the current iterate $\boldsymbol{x}_n$ and majorizing or minorizing $f(\boldsymbol{x})$. As a byproduct of optimizing $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ with respect to $\boldsymbol{x}$, the objective function $f(\boldsymbol{x})$ is sent downhill or uphill, depending on whether the purpose is minimization or maximization. The next iterate $\boldsymbol{x}_{n+1}$ is chosen to optimize the surrogate $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ subject to any relevant constraints. Majorization combines two conditions: the tangency condition $g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) = f(\boldsymbol{x}_n)$ and the domination condition $g(\boldsymbol{x} \mid \boldsymbol{x}_n) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x}$. In minimization these conditions and the definition of $\boldsymbol{x}_{n+1}$ lead to the descent property

$$f(\boldsymbol{x}_{n+1}) \leq g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) \leq g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) = f(\boldsymbol{x}_n).$$

Minorization reverses the domination inequality and produces an ascent algorithm. Under appropriate regularity conditions, an MM algorithm is guaranteed to converge to a stationary point of the objective function [30]. From the perspective of dynamical systems, the objective function serves as a Liapunov function for the algorithm map.

The MM principle simplifies optimization by: (a) separating the variables of a problem, (b) avoiding large matrix inversions, (c) linearizing a problem, (d) restoring symmetry, (e) dealing with equality and inequality constraints gracefully, and (f) turning a nondifferentiable problem into a smooth problem. Choosing a tractable surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ that

Proceedings of the International Congress of Mathematicians, Seoul, 2014

hugs the objective function $f(\boldsymbol{x})$ as tightly as possible requires experience and skill with inequalities. The majorization relation between functions is closed under the formation of sums, nonnegative products, limits, and composition with an increasing function. Hence, it is possible to work piecemeal in majorizing complicated objective functions.

It is impossible to do justice to the complex history of the MM principle in a paragraph. The celebrated EM (expectation-maximization) principle of computational statistics is a special case of the MM principle [33]. Specific MM and EM algorithms appeared years before the principle was well understood [22, 32, 38, 40, 41]. The widely applied projected gradient and proximal gradient algorithms can be motivated from the MM perspective, but the early emphasis on operators and fixed points obscured this distinction. Although Dempster, Laird, and Rubin [15] formally named the EM algorithm, many of their contributions were anticipated by Baum [1] and Sundberg [39]. The MM principle was clearly stated by Ortega and Rheinboldt [36]. de Leeuw [13] is generally credited with recognizing the importance of the principle in practice. The EM algorithm had an immediate and large impact in computational statistics. The more general MM principle was much slower to take hold. The papers [14, 23, 26] by the Dutch school of psychometricians solidified its position. (In this early literature the MM principle is called iterative majorization.) The related Dinklebach [17] maneuver in fractional linear programming also highlighted the importance of the descent property in algorithm construction.

Before moving on, let us record some notational conventions. All vectors and matrices appear in boldface. The * superscript indicates a vector or matrix transpose. The Euclidean norm of a vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|$ and the Frobenius norm of a matrix $\boldsymbol{M}$ by $\|\boldsymbol{M}\|_F$. For a smooth real-valued function $f(\boldsymbol{x})$, we write its gradient (column vector of partial derivatives) as $\nabla f(\boldsymbol{x})$, its first differential (row vector of derivatives) as $df(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^*$, and its second differential (Hessian matrix) as $d^2 f(\boldsymbol{x})$.

2. An adaptive barrier method

In convex programming it simplifies matters notationally to replace a convex inequality constraint $h_j(\boldsymbol{x}) \leq 0$ by the concave constraint $v_j(\boldsymbol{x}) = -h_j(\boldsymbol{x}) \geq 0$. Barrier methods operate on the relative interior of the feasible region where all $v_j(\boldsymbol{x}) > 0$. Adding an appropriate barrier term to the objective function $f(\boldsymbol{x})$ keeps an initially inactive constraint $v_j(\boldsymbol{x})$ inactive throughout an optimization search. If the barrier function is well designed, it should adapt and permit convergence to a feasible point $\boldsymbol{y}$ with one or more inequality constraints active.

We now briefly summarize an adaptive barrier method that does not follow the central path [27]. Because the logarithm of a concave function is concave, the Bregman majorization [7]

$$-\ln v_j(\boldsymbol{x}) + \ln v_j(\boldsymbol{x}_n) + \frac{1}{v_j(\boldsymbol{x}_n)} dv_j(\boldsymbol{x}_n)(\boldsymbol{x} - \boldsymbol{x}_n) \geq 0$$

acts as a convex barrier for a smooth constraint $v_j(\boldsymbol{x}) \geq 0$. To make the barrier adaptive, we scale it by the current value $v_j(\boldsymbol{x}_n)$ of the constraint. These considerations suggest an MM

algorithm based on the surrogate function

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) = f(\boldsymbol{x}) - \rho \sum_{j=1}^{s} v_j(\boldsymbol{x}_n) \ln v_j(\boldsymbol{x}) + \rho \sum_{j=1}^{s} dv_j(\boldsymbol{x}_n)(\boldsymbol{x} - \boldsymbol{x}_n)$$

for s inequality constraints. Minimizing the surrogate subject to relevant linear equality constraints $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ produces the next iterate $\boldsymbol{x}_{n+1}$. The constant ρ determines the tradeoff between keeping the constraints inactive and minimizing $f(\boldsymbol{x})$. One can show that the MM algorithm with exact minimization converges to the constrained minimum of $f(\boldsymbol{x})$ [30].

In practice one step of Newton's method is usually adequate to decrease $f(\boldsymbol{x})$. The first step of Newton's method minimizes the second-order Taylor expansion of $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ around $\boldsymbol{x}_n$ subject to the equality constraints. Given smooth functions, the two differentials

$$\begin{aligned} dg(\boldsymbol{x}_n \mid \boldsymbol{x}_n) =& df(\boldsymbol{x}_n) \\ d^2 g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) =& d^2 f(\boldsymbol{x}_n) - \rho \sum_{j=1}^{s} d^2 v_j(\boldsymbol{x}_n) \\ & + \rho \sum_{j=1}^{s} \frac{1}{v_j(\boldsymbol{x}_n)} \nabla v_j(\boldsymbol{x}_n) dv_j(\boldsymbol{x}_n) \end{aligned} \tag{2.1}$$

are the core ingredients in the quadratic approximation of $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$. Unfortunately, one step of Newton's method is neither guaranteed to decrease $f(\boldsymbol{x})$ nor to respect the nonnegativity constraints.

Example 2.1 (Adaptive Barrier Method for Linear Programming). For instance, the standard form of linear programming requires minimizing a linear function $f(\boldsymbol{x}) = \boldsymbol{c}^* \boldsymbol{x}$ subject to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq \boldsymbol{0}$. The quadratic approximation to the surrogate $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ amounts to

$$\boldsymbol{c}^* \boldsymbol{x}_n + \boldsymbol{c}^* (\boldsymbol{x} - \boldsymbol{x}_n) + \frac{\rho}{2} \sum_{j=1}^{p} \frac{1}{x_{nj}} (x_j - x_{nj})^2.$$

The minimum of this quadratic subject to the linear equality constraints occurs at the point

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \boldsymbol{D}_n^{-1} \boldsymbol{c} + \boldsymbol{D}_n^{-1} \boldsymbol{A}^* (\boldsymbol{A} \boldsymbol{D}_n^{-1} \boldsymbol{A}^*)^{-1} (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{A} \boldsymbol{D}_n^{-1} \boldsymbol{c}).$$

Here $\boldsymbol{D}_n$ is the diagonal matrix with ith diagonal entry ρx_{ni}^{-1}, and the increment $\boldsymbol{x}_{n+1} - \boldsymbol{x}_n$ satisfies the linear equality constraint $\boldsymbol{A}(\boldsymbol{x}_{n+1} - \boldsymbol{x}_n) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_n$.

One can overcome the objections to Newton updates by taking a controlled step along the Newton direction $\boldsymbol{u}_n = \boldsymbol{x}_{n+1} - \boldsymbol{x}_n$. The key is to exploit the theory of self-concordant functions [5, 35]. A thrice differentiable convex function $h(t)$ is said to be self-concordant if it satisfies the inequality

$$|h'''(t)| \leq 2c h''(t)^{3/2}$$

for some constant $c \geq 0$ and all t in the essential domain of $h(t)$. All convex quadratic functions qualify as self-concordant with $c = 0$. The function $h(t) = -\ln(at + b)$ is self-concordant with constant 1. The class of self-concordant functions is closed under sums and

composition with linear functions. A convex function $k(\boldsymbol{x})$ with domain R^p is said to be self-concordant if every slice $h(t) = k(\boldsymbol{x} + t\boldsymbol{u})$ is self-concordant.

Rather than conduct an expensive one-dimensional search along the Newton direction $\boldsymbol{x}_n + t\boldsymbol{u}_n$, one can majorize the surrogate function $h(t) = g(\boldsymbol{x}_n + t\boldsymbol{u}_n \mid \boldsymbol{x}_n)$ along the half-line $t \geq 0$. The clever majorization

$$h(t) \leq h(0) + h'(0)t - \frac{1}{c}h''(0)^{1/2}t - \frac{1}{c^2}\ln[1 - cth''(0)^{1/2}] \tag{2.2}$$

serves the dual purpose of guaranteeing a decrease in $f(\boldsymbol{x})$ and preventing a violation of the inequality constraints [35]. Here c is the self-concordance constant associated with the surrogate. The optimal choice of t reduces to the damped Newton update

$$t = \frac{h'(0)}{h''(0) - ch'(0)h''(0)^{1/2}}. \tag{2.3}$$

The first two derivatives of $h(t)$ are clearly

$$\begin{aligned} h'(0) =& df(\boldsymbol{x}_n)\boldsymbol{u}_n \\ h''(0) =& \boldsymbol{u}_n^* d^2 f(\boldsymbol{x}_n)\boldsymbol{u}_n - \rho \sum_{j=1}^{s} \boldsymbol{u}_n^* d^2 v_j(\boldsymbol{x}_n)\boldsymbol{u}_n \\ &+ \rho \sum_{j=1}^{s} \frac{1}{v_j(\boldsymbol{x}_n)}[dv_j(\boldsymbol{x}_n)\boldsymbol{u}_n]^2. \end{aligned}$$

The first of these derivatives is nonpositive because $\boldsymbol{u}_n$ is a descent direction for $f(\boldsymbol{x})$. The second is generally positive because all of the contributing terms are nonnegative.

	No Safeguard		Self-concordant Safeguard		
Iteration n	$\boldsymbol{c}^*\boldsymbol{x}_n$	$\|\boldsymbol{\Delta}_n\|$	$\boldsymbol{c}^*\boldsymbol{x}_n$	$\|\boldsymbol{\Delta}_n\|$	t_n
1	-1.20000	0.25820	-1.11270	0.14550	0.56351
2	-1.33333	0.17213	-1.20437	0.11835	0.55578
3	-1.41176	0.10125	-1.27682	0.09353	0.55026
4	-1.45455	0.05523	-1.33288	0.07238	0.54630
5	-1.47692	0.02889	-1.37561	0.05517	0.54345
10	-1.49927	0.00094	-1.47289	0.01264	0.53746
15	-1.49998	0.00003	-1.49426	0.00271	0.53622
20	-1.50000	0.00000	-1.49879	0.00057	0.53597
25	-1.50000	0.00000	-1.49975	0.00012	0.53591
30	-1.50000	0.00000	-1.49995	0.00003	0.53590
35	-1.50000	0.00000	-1.49999	0.00001	0.53590
40	-1.50000	0.00000	-1.50000	0.00000	0.53590

Table 2.1. Performance of the adaptive barrier method in linear programming.

When $f(\boldsymbol{x})$ is quadratic and the inequality constraints are affine, detailed calculations show that the surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ is self-concordant with constant

$$c = \frac{1}{\sqrt{\rho \min\{v_1(\boldsymbol{x}_n), \ldots, v_s(\boldsymbol{x}_n)\}}}.$$

Taking the damped Newton's step with step length (2.3) keeps $\boldsymbol{x}_n + t_n \boldsymbol{u}_n$ in the relative interior of the feasible region while decreasing the surrogate and hence the objective function $f(\boldsymbol{x})$. When $f(\boldsymbol{x})$ is not quadratic but can be majorized by a quadratic $q(\boldsymbol{x} \mid \boldsymbol{x}_n)$, one can replace $f(\boldsymbol{x})$ by $q(\boldsymbol{x} \mid \boldsymbol{x}_n)$ in calculating the adaptive-barrier update. The next iterate $\boldsymbol{x}_{n+1}$ retains the descent property.

As a toy example consider the linear programming problem of minimizing $\boldsymbol{c}^* \boldsymbol{x}$ subject to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq \boldsymbol{0}$. Applying the adaptive barrier method to the choices

$$\boldsymbol{A} = \begin{pmatrix} 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{c} = \begin{pmatrix} -1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and to the feasible initial point $\boldsymbol{x}_0 = \frac{1}{3}\boldsymbol{1}$ produces the results displayed in Table 2.1. Not shown is the minimum point $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0, 0)^*$. Columns two and three of the table record the progress of the unadorned adaptive barrier method. The quantity $\|\boldsymbol{\Delta}_n\|$ equals the Euclidean norm of the difference vector $\boldsymbol{\Delta}_n = \boldsymbol{x}_n - \boldsymbol{x}_{n-1}$. Columns four and five repeat this information for the algorithm modified by the self-concordant majorization (2.2). The quantity t_n in column six represents the optimal step length (2.3) in going from $\boldsymbol{x}_{n-1}$ to $\boldsymbol{x}_n$ along the Newton direction $\boldsymbol{u}_{n-1}$. Clearly, there is a price to be paid in implementing a safeguarded Newton step. In practice, this price is well worth paying.

3. Distance majorization

On a Euclidean space, the distance to a closed set S is a Lipschitz function $\operatorname{dist}(\boldsymbol{x}, S)$ with Lipschitz constant 1. If S is also convex, then $\operatorname{dist}(\boldsymbol{x}, S)$ is a convex function. Projection onto S is intimately tied to $\operatorname{dist}(\boldsymbol{x}, S)$. Unless S is convex, the projection operator $P_S(\boldsymbol{x})$ is multi-valued for at least one argument $\boldsymbol{x}$. Fortunately, it is possible to majorize $\operatorname{dist}(\boldsymbol{x}, S)$ at $\boldsymbol{x}_n$ by $\|\boldsymbol{x} - P_S(\boldsymbol{x}_n)\|$. This simple observation is the key to the proximal distance algorithm to be discussed later. In the meantime, let us show how to derive two feasibility algorithms by distance majorization [9]. Let $S_1, \ldots, S_m$ be closed sets. The method of averaged projections attempts to find a point in their intersection $S = \cap_{j=1}^m S_j$. To derive the algorithm, consider the convex combination

$$f(\boldsymbol{x}) = \sum_{j=1}^m \alpha_j \operatorname{dist}(\boldsymbol{x}, S_j)^2$$

of squared distance functions. Obviously, $f(\boldsymbol{x})$ vanishes precisely on S when all $\alpha_j > 0$. The majorization

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) = \sum_{j=1}^m \alpha_j \|\boldsymbol{x} - P_{S_j}(\boldsymbol{x}_n)\|^2$$

of $f(\boldsymbol{x})$ is easily minimized. The minimum point of $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$,

$$\boldsymbol{x}_{n+1} = \sum_{j=1}^{m} \alpha_j P_{S_j}(\boldsymbol{x}_n),$$

defines the averaged operator. The MM principle guarantees that $\boldsymbol{x}_{n+1}$ decreases the objective function.

Von Neumann's method of alternating projections can also be derived from this perspective. For two sets S_1 and S_2, consider the problem of minimizing the objective function $f(\boldsymbol{x}) = \operatorname{dist}(\boldsymbol{x}, S_2)^2$ subject to the constraint $\boldsymbol{x} \in S_1$. The function

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) = \|\boldsymbol{x} - P_{S_2}(\boldsymbol{x}_n)\|^2$$

majorizes $f(\boldsymbol{x})$. Indeed, the domination condition $g(\boldsymbol{x} \mid \boldsymbol{x}_n) \geq f(\boldsymbol{x})$ holds because $P_{S_2}(\boldsymbol{x}_n)$ belongs to S_2; the tangency condition $g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) = f(\boldsymbol{x}_n)$ holds because $P_{S_2}(\boldsymbol{x}_n)$ is the closest point in S_2 to $\boldsymbol{x}_n$. The surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ is minimized subject to the constraint by taking $\boldsymbol{x}_{n+1} = P_{S_1} \circ P_{S_2}(\boldsymbol{x}_n)$. The MM principle again ensures that $\boldsymbol{x}_{n+1}$ decreases the objective function. When the two sets intersect, the least distance of 0 is achieved at any point in the intersection. One can extend this derivation to three sets by minimizing the objective function $f(\boldsymbol{x}) = \operatorname{dist}(\boldsymbol{x}, S_2)^2 + \operatorname{dist}(\boldsymbol{x}, S_3)^2$ subject to $\boldsymbol{x} \in S_1$. The surrogate

$$\begin{aligned} g(\boldsymbol{x} \mid \boldsymbol{x}_n) &= \|\boldsymbol{x} - P_{S_2}(\boldsymbol{x}_n)\|^2 + \|\boldsymbol{x} - P_{S_3}(\boldsymbol{x}_n)\|^2 \\ &= 2\left\|\boldsymbol{x} - \frac{1}{2}[P_{S_2}(\boldsymbol{x}_n) + P_{S_3}(\boldsymbol{x}_n)]\right\|^2 + c_n \end{aligned}$$

relies on an irrelevant constant c_n. The closest point in S_1 is

$$\boldsymbol{x}_{n+1} = P_{S_1}\left\{\frac{1}{2}[P_{S_2}(\boldsymbol{x}_n) + P_{S_3}(\boldsymbol{x}_n)]\right\}.$$

This construction clearly generalizes to more than three sets.

4. The proximal distance method

We now turn to an exact penalty method that applies to nonsmooth functions. Clarke's exact penalty method [10] turns the constrained problem of minimizing a function $f(\boldsymbol{y})$ over a closed set S into the unconstrained problem of minimizing the function $f(\boldsymbol{y}) + \rho \operatorname{dist}(\boldsymbol{y}, S)$ for ρ sufficiently large. Here is a precise statement of a generalization of Clarke's result [6, 10, 11].

Proposition 4.1. *Suppose $f(\boldsymbol{y})$ achieves a local minimum on S at the point $\boldsymbol{x}$. Let $\phi_S(\boldsymbol{y})$ denote a function that vanishes on S and satisfies $\phi_S(\boldsymbol{y}) \geq c \operatorname{dist}(\boldsymbol{y}, S)$ for all $\boldsymbol{x}$ and some positive constant c. If $f(\boldsymbol{y})$ is locally Lipschitz around $\boldsymbol{x}$ with constant L, then for every $\rho \geq c^{-1}L$, $F_\rho(\boldsymbol{y}) = f(\boldsymbol{y}) + \rho\phi_S(\boldsymbol{y})$ achieves a local unconstrained minimum at $\boldsymbol{x}$.*

Classically the choice $\phi_S(\boldsymbol{x}) = \operatorname{dist}(\boldsymbol{x}, S)$ was preferred. For affine equality constraints $g_i(\boldsymbol{x}) = 0$ and affine inequality constraints $h_j(\boldsymbol{x}) \leq 0$, Hoffman's bound

$$\operatorname{dist}(\boldsymbol{y}, S) \leq \tau\rho \left\| \begin{matrix} G(\boldsymbol{y}) \\ H(\boldsymbol{y})_+ \end{matrix} \right\|$$

applies, where τ is some positive constant, S is the feasible set where $G(\boldsymbol{y}) = \mathbf{0}$, and $H(\boldsymbol{y})_+ \leq \mathbf{0}$ [24]. The vector $H(\boldsymbol{y})_+$ has components $h_j(\boldsymbol{x})_+ = \max\{h_j(\boldsymbol{y}), 0\}$. When S is the intersection of several closed sets $S_1, \ldots, S_m$, the alternative

$$\phi_S(\boldsymbol{y}) = \sqrt{\sum_{i=1}^{m} \operatorname{dist}(\boldsymbol{y}, C_i)^2} \tag{4.1}$$

is attractive. The next proposition gives sufficient conditions under which the crucial bound $\phi_S(\boldsymbol{y}) \geq c \operatorname{dist}(\boldsymbol{y}, S)$ is valid for the function (4.1).

Proposition 4.2. *Suppose $S_1, \ldots, S_m$ are closed convex sets in R^p with the first j sets polyhedral. Assume further that the intersection*

$$S = (\cap_{i=1}^{j} S_i) \cap (\cap_{i=j+1}^{m} \operatorname{ri} S_i)$$

is nonempty and bounded. Then there exists a constant $\tau > 0$ such that

$$\operatorname{dist}(\boldsymbol{x}, S) \leq \tau \sum_{i=1}^{m} \operatorname{dist}(\boldsymbol{x}, S_i) \leq \tau\sqrt{m} \sqrt{\sum_{i=1}^{m} \operatorname{dist}(\boldsymbol{x}, S_i)^2}$$

for all $\boldsymbol{x}$. The sets $S_1, \ldots, S_m$ are said to be linearly regular.

Proof. See the references [2, 16] for all details. A polyhedral set is the nonempty intersection of a finite number of half-spaces. The operator $\operatorname{ri} K$ forms the relative interior of the convex set K, namely, the interior of K relative to the affine hull of K. When K is nonempty, its relative interior is nonempty and generates the same affine hull as K itself. □

In general, we will require $f(\boldsymbol{x})$ and $\phi_S(\boldsymbol{x})$ to be continuous functions and the sum $F_\rho(\boldsymbol{y}) = f(\boldsymbol{y}) + \rho\phi_S(\boldsymbol{y})$ to be coercive for some value $\rho = \rho_0$. It then follows that $F_\rho(\boldsymbol{y})$ is coercive and attains its minimum for all $\rho \geq \rho_0$. One can prove a partial converse to Clarke's theorem [11, 12]. This requires the enlarged set $S_\epsilon = \{\boldsymbol{x} : \phi_S(\boldsymbol{x}) < \epsilon\}$ of points lying close to S as measured by $\phi_S(\boldsymbol{x})$.

Proposition 4.3. *Suppose that $f(\boldsymbol{y})$ is Lipschitz on S_ϵ for some $\epsilon > 0$. Then under the stated assumptions on $f(\boldsymbol{x})$ and $\phi_S(\boldsymbol{x})$, a global minimizer of $F_\rho(\boldsymbol{y})$ is a constrained minimizer of $f(\boldsymbol{y})$ for all sufficiently large ρ.*

When the constraint set S is compact and $f(\boldsymbol{y})$ has a continuously varying local Lipschitz constant, the hypotheses of Proposition 4.3 are fulfilled. This is the case, for instance, when $f(\boldsymbol{y})$ is continuously differentiable. With this background on the exact penalty method in mind, we now sketch an approximate MM algorithm for convex programming that is motivated by distance majorization. This algorithm is designed to exploit set projections and proximal maps. The proximal map $\operatorname{prox}_h(\boldsymbol{y})$ associated with a convex function $h(\boldsymbol{x})$ satisfies

$$\operatorname{prox}_h(\boldsymbol{y}) = \operatorname{argmin}_{\boldsymbol{x}} \left[h(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \right].$$

A huge literature and software base exist for computing projections and proximal maps [3].

Since the function $\operatorname{dist}(\boldsymbol{x}, S)$ is merely continuous, we advocate approximating it by the differentiable function

$$\operatorname{dist}_\epsilon(\boldsymbol{x}, S) = \sqrt{\operatorname{dist}(\boldsymbol{x}, S)^2 + \epsilon}$$

for $\epsilon > 0$ small. The composite function $\operatorname{dist}_\epsilon(\boldsymbol{x}, S)$ is convex when S is convex because the function $\sqrt{t^2 + \epsilon}$ is increasing and convex on $[0, \infty)$. Instead of minimizing $f(\boldsymbol{x}) + \rho \operatorname{dist}(\boldsymbol{x}, S)$, we suggest minimizing the differentiable convex function $f(\boldsymbol{x}) + \rho \operatorname{dist}_\epsilon(\boldsymbol{x}, S)$ by an MM algorithm. Regardless of whether S is convex, the majorization

$$\operatorname{dist}_\epsilon(\boldsymbol{x}, S) \leq \sqrt{\|\boldsymbol{x} - P_S(\boldsymbol{x}_n)\|^2 + \epsilon} \tag{4.2}$$

holds. If S is nonconvex, there may be a multiplicity of closest points, and one must choose a representative of the set $P_S(\boldsymbol{x}_n)$. In any event one can invoke the univariate majorization

$$\sqrt{t} \geq \sqrt{t_n} + \frac{t - t_n}{2\sqrt{t_n}} \tag{4.3}$$

of the concave function $\sqrt{t}$ on the interval $t > 0$ and majorize the majorization (4.2) by

$$\sqrt{\|\boldsymbol{x} - P_S(\boldsymbol{x}_n)\|^2 + \epsilon} \leq \frac{1}{2\sqrt{\|\boldsymbol{x}_n - P_S(\boldsymbol{x}_n)\|^2 + \epsilon}} \|\boldsymbol{x} - P_S(\boldsymbol{x}_n)\|^2 + c_n$$

for some irrelevant constant c_n. The second step of our proposed MM algorithm consists of minimizing the surrogate function

$$\begin{aligned} g(\boldsymbol{x} \mid \boldsymbol{x}_n) &= f(\boldsymbol{x}) + \frac{w_n}{2} \|\boldsymbol{x} - P_S(\boldsymbol{x}_n)\|^2 \\ w_n &= \frac{\rho}{\sqrt{\|\boldsymbol{x}_n - P_S(\boldsymbol{x}_n)\|^2 + \epsilon}}. \end{aligned}$$

The corresponding proximal map drives $f(\boldsymbol{x}) + \rho \operatorname{dist}_\epsilon(\boldsymbol{x}, S)$ downhill. Under the more general exact penalty (4.1), the surrogate function depends on a sum of spherical quadratics rather than a single spherical quadratic.

It is possible to project onto a variety of closed nonconvex sets. For example, if S is the set of integers, then projection amounts to rounding. An ambiguous point $n + \frac{1}{2}$ can be projected to either n or $n + 1$. Projection onto a finite set simply tests each point separately. Projection onto a Cartesian product is achieved via the Cartesian product of the projections. One can also project onto many continuous sets of interest. For example, to project onto the closed set of points having at most k nonzero coordinates, one zeros out all but the k largest coordinates in absolute value. Projection onto the sphere of center $\boldsymbol{z}$ and radius r takes $\boldsymbol{y} \neq \boldsymbol{z}$ into the point $\boldsymbol{z} + \frac{r}{\|\boldsymbol{y} - \boldsymbol{z}\|}(\boldsymbol{y} - \boldsymbol{z})$. All points of the sphere are equidistant from its center.

By definition the update $\boldsymbol{x}_{n+1} = \operatorname{prox}_{w_n^{-1} f}[P_S(\boldsymbol{x}_n)]$ minimizes $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$. We will refer to this MM algorithm as the **proximal distance algorithm**. It enjoys several virtues. First, it allows one to exploit the extensive body of results on proximal maps and projections. Second, it does not demand that the constraint set S be convex. Third, it does not require the objective function $f(\boldsymbol{x})$ to be convex or smooth. Finally, the minimum values and minimum points of the functions $f(\boldsymbol{x}) + \rho \operatorname{dist}(\boldsymbol{x}, S)$ and $f(\boldsymbol{x}) + \rho \operatorname{dist}_\epsilon(\boldsymbol{x}, S)$ are close when $\epsilon > 0$ is small.

In implementing the proximal distance algorithm, the constants L and ϵ must specified. For many norms the Lipschitz constant L is known. For a differentiable function $f(\boldsymbol{x})$, the mean value inequality suggests taking L equal to the maximal value of $\|\nabla f(\boldsymbol{x})\|$ in a neighborhood of the optimal point. In specific problems a priori bounds can be derived. If no such prior bound is known, then one has to guess an appropriate ρ and see if it leads to a constrained minimum. If not, ρ should be systematically increased until a constrained minimum is reached. Even with a justifiable bound, it is prudent to start ρ well below its intended upper bound to emphasize minimization of the loss function in early iterations. Experience shows that gradually decreasing ϵ is also a good tactic; otherwise, one again runs the risk of putting too much early stress on satisfying the constraints. In practice the sequences $\rho_n = \min\{\alpha^n \rho_0, \rho_{\max}\}$ and $\epsilon_n = \max\{\beta^{-n}\epsilon_0, \epsilon_{\min}\}$ work well for α and β slightly larger than 1, say 1.2, and $\rho_0 = \epsilon_0 = 1$. On many problems more aggressive choices of α and β are possible. The values of $\rho_{\max}$ and $\epsilon_{\min}$ are problem specific, but taking $\rho_{\max}$ substantially greater than a known Lipschitz constant slows convergence. Taking $\epsilon_{\min}$ too large leads to a poor approximate solution.

5. Sample problems

We now explore some typical applications of the proximal distance algorithm. In all cases we are able to establish local Lipschitz constants. Comparisons with standard optimization software serve as performance benchmarks.

Example 5.1 (Projection onto an Intersection of Closed Convex Sets). Let $S_1, \ldots, S_k$ be closed convex sets, and assume that projection onto each S_j is straightforward. Dykstra's algorithm [16, 18] is designed to find the projection of an external point $\boldsymbol{y}$ onto $S = \cap_{j=1}^k S_j$. The proximal distance algorithm provides an alternative based on the convex function

$$f(\boldsymbol{x}) = \sqrt{\|\boldsymbol{x} - \boldsymbol{y}\|^2 + \delta}$$

for δ positive, say $\delta = 1$. The choice $f(\boldsymbol{x})$ is preferable to the obvious choice $\|\boldsymbol{x} - \boldsymbol{y}\|^2$ because $f(\boldsymbol{x})$ is Lipschitz with Lipschitz constant 1. In the proximal distance algorithm, we take

$$\phi_S(\boldsymbol{x}) = \sqrt{\sum_{j=1}^k \operatorname{dist}(\boldsymbol{x}, S_j)^2}$$

and minimize the surrogate function

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) = f(\boldsymbol{x}) + \frac{w_n}{2} \sum_{j=1}^k \|\boldsymbol{x} - \boldsymbol{p}_{nj}\|^2 = f(\boldsymbol{x}) + \frac{k w_n}{2} \|\boldsymbol{x} - \bar{\boldsymbol{p}}_n\|^2 + c_n,$$

where $\boldsymbol{p}_{nj}$ is the projection of $\boldsymbol{x}_n$ onto S_j, $\bar{\boldsymbol{p}}_n$ is the average of the projections $\boldsymbol{p}_{nj}$, c_n is an irrelevant constant, and

$$w_n = \frac{\rho}{\sqrt{\sum_{j=1}^k \|\boldsymbol{x}_n - \boldsymbol{p}_{nj}\|^2 + \epsilon}}.$$

After rearrangement, the stationarity condition for optimality reads

$$\boldsymbol{x} = (1-\alpha)\boldsymbol{y} + \alpha\bar{\boldsymbol{p}}_n, \quad \alpha = \frac{kw_n}{\frac{1}{\sqrt{\|\boldsymbol{x}-\boldsymbol{y}\|^2+\delta}} + kw_n}.$$

In other words, $\boldsymbol{x}_{n+1}$ is a convex combination of $\boldsymbol{y}$ and $\bar{\boldsymbol{p}}_n$.

	Dykstra		Proximal Distance	
Iteration n	x_{n1}	x_{n2}	x_{n1}	x_{n2}
0	-1.00000	2.00000	-1.00000	2.00000
1	-0.44721	0.89443	-0.44024	1.60145
2	0.00000	0.89443	-0.25794	1.38652
3	-0.26640	0.96386	-0.16711	1.25271
4	0.00000	0.96386	-0.11345	1.16647
5	-0.14175	0.98990	-0.07891	1.11036
10	0.00000	0.99934	-0.01410	1.01576
15	-0.00454	0.99999	-0.00250	1.00257
20	0.00000	1.00000	-0.00044	1.00044
25	-0.00014	1.00000	-0.00008	1.00008
30	0.00000	1.00000	-0.00001	1.00001
35	0.00000	1.00000	0.00000	1.00000

Table 5.1. Dykstra's algorithm versus the proximal distance algorithm.

To calculate the optimal coefficient α, we minimize the convex surrogate

$$h(\alpha) = g[(1-\alpha)\boldsymbol{y} + \alpha\bar{\boldsymbol{p}}_n \mid \boldsymbol{x}_n] = \sqrt{\alpha^2 d^2 + \delta} + \frac{kw_n}{2}(1-\alpha)^2 d^2 + c_n$$

for $d = \|\boldsymbol{y} - \bar{\boldsymbol{p}}_n\|$. Its derivative

$$h'(\alpha) = \frac{\alpha d^2}{\sqrt{\alpha^2 d^2 + \delta}} - kw_n(1-\alpha)d^2$$

satisfies $h'(0) < 0$ and $h'(1) > 0$ and possesses a unique root on the open interval $(0,1)$. This root can be easily computed by bisection or Newton's method.

Table 5.1 compares Dykstra's algorithm and the proximal distance algorithm on a simple planar example. Here S_1 is the closed unit ball in R^2, and S_2 is the closed halfspace with $x_1 \geq 0$. The intersection S reduces to the right half ball centered at the origin. The table records the iterates of the two algorithms from the starting point $\boldsymbol{x}_0 = (-1,2)^*$ until their eventual convergence to the geometrically obvious solution $(0,1)^*$. In the proximal distance method we set $\rho_n = 2$ and aggressively $\epsilon_n = 4^{-n}$. The two algorithms exhibit similar performance but take rather different trajectories.

Example 5.2 (Linear Programming). The standard version of linear programming minimizes $f(\boldsymbol{x}) = \boldsymbol{c}^*\boldsymbol{x}$ subject to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq \boldsymbol{0}$. The norm $\|\boldsymbol{c}\|$ serves as a Lipschitz constant for $f(\boldsymbol{x})$. Projection onto the affine space $S = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ is achieved via

$$P_S(\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}).$$

Variables	Constraints	MM	CVX	MATLAB	YALMIP
2	4	0.010	0.100	0.005	0.088
4	8	0.007	0.070	0.005	0.117
8	4	0.012	0.080	0.004	0.141
16	8	0.008	0.080	0.005	0.213
32	64	0.012	0.090	0.005	0.161
64	128	0.016	0.110	0.010	0.132
128	256	0.026	0.160	0.033	0.193
256	512	0.055	0.370	0.187	0.320
512	256	0.214	1.210	1.358	1.656
1024	2048	1.302	11.920	10.883	12.129
2048	4016	8.721	85.330	78.114	79.630
4096	8192	59.044	881.970	562.648	593.823

Table 5.2. Computation times in seconds for various linear programs.

Computing the pseudoinverse $\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}$ once and storing it improves performance. Projection onto $\mathsf{R}^d_+ = \{\boldsymbol{x} \geq \boldsymbol{0}\}$ is trivial. The proximal distance algorithm minimizes the criterion

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) = \boldsymbol{c}^*\boldsymbol{x} + \frac{w_n}{2}\left(\|\boldsymbol{x} - P_S(\boldsymbol{x}_n)\|^2 + \|\boldsymbol{x} - P_{\mathsf{R}^d_+}(\boldsymbol{x}_n)\|^2\right).$$

for the weight

$$w_n = \frac{\rho}{\sqrt{\|\boldsymbol{x}_n - P_S(\boldsymbol{x}_n)\|^2 + \|\boldsymbol{x}_n - P_{\mathsf{R}^d_+}(\boldsymbol{x}_n)\|^2 + \epsilon}}.$$

The update

$$\boldsymbol{x}_{n+1} = -\frac{1}{2w_n}\boldsymbol{c} + \frac{1}{2}P_S(\boldsymbol{x}_n) + \frac{1}{2}P_{\mathsf{R}^d_+}(\boldsymbol{x}_n)$$

is straightforward to derive and simple to implement.

For obscure reasons the MM proximal algorithm exhibits better performance on overdetermined problems than on underdetermined ones. We handle an underdetermined problem by solving its overdetermined dual problem regardless of whether the dual requires more variables. Because the dual linear program minimizes $\boldsymbol{b}^*\boldsymbol{y}$ subject to $\boldsymbol{A}^*\boldsymbol{y} = \boldsymbol{c}$ and $\boldsymbol{y} \geq \boldsymbol{0}$, the Lipschitz bound for the dual is $\|\boldsymbol{b}\|$. We compared a MATLAB implementation of the MM algorithm to CVX [21] with the SeDuMi solver, YALMIP [31] with the MOSEK solver, and MATLAB's `linprog` routine. For overdetermined problems, we asked the MM algorithm, MATLAB, and YALMIP to solve the primal problem. For underdetermined problems, we reversed this strategy. CVX always solved the primal problem. Our test problems involve d variables and $2d$ constraints or vice versa. We filled $\boldsymbol{A}$ with standard normal deviates and two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ with standard uniform deviates and set $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{v}$ and $\boldsymbol{c} = \boldsymbol{A}^*\boldsymbol{u}$, ensuring both primal and dual feasibility. Our gentle tuning constant schedules $\rho_n = \min\{1.2^n, 2L\}$ and $\epsilon_n = \max\{1.2^n, 10^{-15}\}$ required either the Lipschitz bound $L = \|\boldsymbol{c}\|$ or $L = \|\boldsymbol{b}\|$ as just noted. For each problem summarized in Table 5.2, the four converged solutions agree to at least five digits. The table demonstrates the substantial speed advantage of the MM algorithm on moderately large problems.

Example 5.3 (Binary Piecewise-Linear Functions). The problem of minimizing the binary piecewise-linear function

$$f(\boldsymbol{x}) = \sum_{i<j} w_{ij}|x_i - x_j| + \boldsymbol{b}^*\boldsymbol{x}$$

subject to $\boldsymbol{x} \in \{0,1\}^d$ and nonnegative weights w_{ij} is a typical discrete optimization problem with applications in graph cuts. If we invoke the majorization

$$|x_i - x_j| \le \left|x_i - \frac{x_{ni} + x_{nj}}{2}\right| + \left|x_j - \frac{x_{ni} + x_{nj}}{2}\right|$$

prior to applying the proximal operator, then the proximal distance algorithm separates the parameters. Parameter separation promotes parallelization and benefits from a fast algorithm for computing proximal maps in one dimension. The one-dimensional algorithm is similar to but faster than bisection [37]. Finally, the objective function is Lipschitz with the explicit constant

$$L = \sum_i \sqrt{\sum_{j\neq i} w_{ij}^2} + \|\boldsymbol{b}\|. \tag{5.1}$$

This assertion follows from the simple bound

$$\begin{aligned} |f(\boldsymbol{x}) - f(\boldsymbol{y})| &\le \sum_i \sum_{j\neq i} w_{ij}|x_j - y_j| + |\boldsymbol{b}^*(\boldsymbol{x} - \boldsymbol{y})| \\ &\le \sum_i \sqrt{\sum_{j\neq i} w_{ij}^2} \cdot \|\boldsymbol{x} - \boldsymbol{y}\| + \|\boldsymbol{b}\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\| \end{aligned}$$

under the symmetry convention $w_{ij} = w_{ji}$.

Table 5.3 displays the numerical results for a few typical examples. For each dimension d we filled $\boldsymbol{b}$ with standard normal deviates and the upper triangle of the weight matrix $\boldsymbol{W}$ with the absolute values of such deviates. The lower triangle of $\boldsymbol{W}$ was determined by symmetry. Small values of $\boldsymbol{b}$ often lead to degenerate solutions $\boldsymbol{x}$ with all entries 0 or 1. To avoid this possibility, we multiplied each entry of $\boldsymbol{b}$ by d. In the graph cut context, a degenerate solution corresponds to no cuts at all or a completely cut graph. These examples depend on the schedules $\rho_n = \min\{1.2^n, L\}$ and $\epsilon_n = \max\{1.2^{-n}, 10^{-15}\}$ for the two tuning constants and the local Lipschitz constant (5.1).

Although the MM proximal distance algorithm makes good progress towards the minimum in the first 100 iterations, it sometimes hovers around its limit without fully converging. This translates into fickle compute times, and for this reason we capped the number of MM iterations at 200. For small dimensions MM can be much slower than CVX. Fortunately, the performance of the MM algorithm improves markedly as d increases. In all runs the two algorithms reach the same solution after rounding components to the nearest integer. MM also requires much less storage than CVX. Asterisks appear in the table where CVX demanded more memory than our laptop computer could deliver.

Example 5.4 (Nonnegative Quadratic Programming). The proximal distance algorithm is applicable in minimizing a convex quadratic $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x}$ subject to the constraint

	CPU times		
Dimension	MM	CVX	Iterations
2	0.038	0.080	9
4	0.052	0.060	18
8	2.007	0.050	200
16	2.416	0.100	200
32	2.251	0.130	200
64	4.134	0.400	200
128	0.212	2.980	32
256	0.868	62.63	200
512	68.27	1534	200
1024	526.6	*	200
2048	127.2	*	200
4096	547.4	*	200

Table 5.3. CPU times in seconds and MM iterations until convergence for binary piecewise linear functions. Asterisks denote computer runs exceeding computer memory limits. Iterations were capped at 200.

$\boldsymbol{x} \geq \boldsymbol{0}$. In this nonnegative quadratic programming program, let $\boldsymbol{y}_n$ be the projection of the current iterate $\boldsymbol{x}_n$ onto $S = \mathsf{R}^d_+$. If we define the weight

$$w_n = \frac{\rho}{\sqrt{\|\boldsymbol{x}_n - \boldsymbol{y}_n\|^2 + \epsilon}},$$

then the next iterate can be expressed as

$$\boldsymbol{x}_{n+1} = (\boldsymbol{A} + w_n \boldsymbol{I})^{-1}(w_n \boldsymbol{y}_n - \boldsymbol{b}).$$

The multiple matrix inversions implied by the update can be avoided by extracting and caching the spectral decomposition $\boldsymbol{U}^* \boldsymbol{D} \boldsymbol{U}$ of $\boldsymbol{A}$ at the start of the algorithm. The inverse $(\boldsymbol{A} + w_n \boldsymbol{I})^{-1}$ then reduces to $\boldsymbol{U}^*(\boldsymbol{D} + w_n \boldsymbol{I})^{-1}\boldsymbol{U}$. The diagonal matrix $\boldsymbol{D} + w_n \boldsymbol{I}$ is obviously trivial to invert. The remaining operations in computing $\boldsymbol{x}_{n+1}$ collapse to matrix times vector multiplications. Nonnegative least squares is a special case of nonnegative quadratic programming.

One can estimate an approximate Lipschitz constant for this problem. Note that $f(\boldsymbol{0}) = 0$ and that

$$f(\boldsymbol{x}) \geq \frac{1}{2}\lambda_{\min}\|\boldsymbol{x}\|^2 - \|\boldsymbol{b}\| \cdot \|\boldsymbol{x}\|,$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\boldsymbol{A}$. It follows that any point $\boldsymbol{x}$ with $\|\boldsymbol{x}\| > \frac{2}{\lambda_{\min}}\|\boldsymbol{b}\|$ cannot minimize $f(\boldsymbol{x})$ subject to the nonnegativity constraint. On the other hand, the gradient of $f(\boldsymbol{x})$ satisfies

$$\|\nabla f(\boldsymbol{x})\| \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\| + \|\boldsymbol{b}\| \leq \lambda_{\max}\|\boldsymbol{x}\| + \|\boldsymbol{b}\|.$$

In view of the mean-value inequality, these bounds suggest that

$$L = \left(\frac{2\lambda_{\max}}{\lambda_{\min}} + 1\right)\|\boldsymbol{b}\| = [2\,\mathrm{cond}_2(\boldsymbol{A}) + 1]\,\|\boldsymbol{b}\|$$

	CPU times				Optima			
d	MM	CV	MA	YA	MM	CV	MA	YA
8	0.97	0.23	0.01	0.13	-0.0172	-0.0172	-0.0172	-0.0172
16	0.50	0.24	0.01	0.11	-1.1295	-1.1295	-1.1295	-1.1295
32	0.50	0.24	0.01	0.14	-1.3811	-1.3811	-1.3811	-1.3811
64	0.57	0.28	0.01	0.13	-0.5641	-0.5641	-0.5641	-0.5641
128	0.79	0.36	0.02	0.14	-0.7018	-0.7018	-0.7018	-0.7018
256	1.66	0.65	0.06	0.22	-0.6890	-0.6890	-0.6890	-0.6890
512	5.61	2.95	0.26	0.73	-0.5971	-0.5968	-0.5970	-0.5970
1024	32.69	21.90	1.32	2.91	-0.4944	-0.4940	-0.4944	-0.4944
2048	156.7	178.8	8.96	15.89	-0.4514	-0.4505	-0.4512	-0.4512
4096	695.1	1551	57.73	91.54	-0.4690	-0.4678	-0.4686	-0.4686

Table 5.4. CPU times in seconds and optima for the nonnegative quadratic program. Abbreviations: d stands for problem dimension, MM for the proximal distance algorithm, CV for CVX, MA for MATLAB's `quadprog`, and YA for YALMIP.

provides an approximate Lipschitz constant for $f(\boldsymbol{x})$ on the region harboring the minimum point. This bound on ρ is usually too large. One remedy is to multiply the bound by a deflation factor such as 0.1. Another remedy is to replace the covariance $\boldsymbol{A}$ by the corresponding correlation matrix. Thus, one solves the problem for the preconditioned matrix $\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{D}^{-1}$, where $\boldsymbol{D}$ is the diagonal matrix whose entries are the square roots of the corresponding diagonal entries of $\boldsymbol{A}$. The transformed parameters $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$ obey the same nonnegativity constraints as $\boldsymbol{x}$.

For testing purposes we filled a $d \times d$ matrix $\boldsymbol{M}$ with independent standard normal deviates and set $\boldsymbol{A} = \boldsymbol{M}^*\boldsymbol{M} + \boldsymbol{I}$. Addition of the identity matrix avoids ill conditioning. We also filled the vector $\boldsymbol{b}$ with independent standard normal deviates. Our gentle tuning constant schedule $\epsilon_n = \max\{1.005^{-n}, 10^{-15}\}$ and $\rho_n = \min\{1.005^n, 0.1 \times L\}$ adjusts ρ and ϵ so slowly that their limits are not actually met in practice. In any event L is the a priori bound for the correlation matrix derived from $\boldsymbol{A}$. Table 5.4 compares the performance of the MM proximal distance algorithm to MATLAB's `quadprog`, CVX with the SDPT3 solver, and YALMIP with the MOSEK solver. MATLAB's `quadprog` is clearly the fastest of the four tested methods on these problems. The relative speed of the MM algorithm improves as the problem dimension d increases.

Example 5.5 (Linear Regression under an ℓ_0 Constraint). In this example the objective function is the sum of squares $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, where $\boldsymbol{y}$ is the response vector, $\boldsymbol{X}$ is the design matrix, and $\boldsymbol{\beta}$ is the vector of regression coefficients. The constraint set S_k^d consists of those $\boldsymbol{\beta}$ with at most k nonzero entries. Projection onto the closed but nonconvex set S_k^d is achieved by zeroing out all but the k largest coordinates in absolute value. These coordinates will be unique except in the rare circumstances of ties. The proximal distance algorithm for this problem coincides with that of the previous problem if we substitute $\boldsymbol{X}^*\boldsymbol{X}$ for $\boldsymbol{A}$, $-\boldsymbol{X}^*\boldsymbol{y}$ for $\boldsymbol{b}$, $\boldsymbol{\beta}$ for $\boldsymbol{x}$, and the projection operator $P_{S_k^d}$ for $P_{\mathbb{R}_+^d}$. Better accuracy can be maintained if the MM update exploits the singular value decomposition of $\boldsymbol{X}$ in forming the spectral decomposition of $\boldsymbol{X}^*\boldsymbol{X}$. Although the proximal distance algorithm carries no absolute guarantee of finding the optimal set of k regression coefficients, it is far more efficient than sifting through all $\binom{d}{k}$ sets of size k. The alternative of lasso-guided

m	n	df	tp_1	tp_2	λ	L_1	L_1/L_2	T_1	T_1/T_2
256	128	10	5.97	3.32	0.143	248.763	0.868	0.603	8.098
128	256	10	3.83	1.91	0.214	106.234	0.744	0.999	10.254
512	256	10	6.51	2.88	0.119	506.570	0.900	0.907	6.262
256	512	10	4.50	1.82	0.172	241.678	0.835	1.743	8.687
1024	512	10	7.80	5.25	0.101	1029.333	0.921	2.597	5.057
512	1024	10	5.54	2.58	0.138	507.451	0.881	8.235	13.532
2048	1024	10	8.98	8.49	0.080	2047.098	0.945	15.460	8.858
1024	2048	10	6.80	2.93	0.110	1044.640	0.916	34.997	18.433
4096	2048	10	9.75	9.90	0.060	4086.886	0.966	89.684	10.956
2048	4096	10	8.36	6.60	0.086	2045.645	0.942	166.386	25.821

Table 5.5. Numerical experiments comparing MM to MATLAB's `lasso`. Each row presents averages over 100 independent simulations. Abbreviations: m the number of cases, n the number of predictors, df the number of actual predictors in the generating model, tp_1 the number of true predictors selected by MM, tp_2 the number of true predictors selected by the lasso, λ the regularization parameter at the lasso optimal loss, L_1 the optimal loss from MM, L_1/L_2 the ratio of L_1 to the optimal lasso loss, T_1 the total computation time in seconds for MM, and T_1/T_2 the ratio of T_1 to the total computation time of the lasso.

model selection must contend with strong shrinkage and a surplus of false positives.

Table 5.5 compares the MM proximal distance algorithm to MATLAB's `lasso` function. In simulating data, we filled $\boldsymbol{X}$ with standard normal deviates, set all components of $\boldsymbol{\beta}$ to 0 except for $\beta_i = 1/i$ for $1 \leq i \leq 10$, and added a vector of standard normal deviates to $\boldsymbol{X}\boldsymbol{\beta}$ to determine $\boldsymbol{y}$. For a given choice of m and n we ran each experiment 100 times and averaged the results. The table demonstrates the superior speed of the lasso and the superior accuracy of the MM algorithm as measured by optimal loss and model selection.

Example 5.6 (Matrix Completion). Let $\boldsymbol{Y} = (y_{ij})$ denote a partially observed $p \times q$ matrix and Δ the set of index pairs (i, j) with y_{ij} observed. Matrix completion [8] imputes the missing entries by approximating $\boldsymbol{Y}$ with a low rank matrix $\boldsymbol{X}$. Imputation relies on the singular value decomposition

$$\boldsymbol{X} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^t, \tag{5.2}$$

where r is the rank of $\boldsymbol{X}$, the nonnegative singular values σ_i are presented in decreasing order, the left singular vectors $\boldsymbol{u}_i$ are orthonormal, and the right singular vectors $\boldsymbol{v}_i$ are also orthonormal [20]. The set R_k of $p \times q$ matrices of rank k or less is closed. Projection onto R_k is accomplished by truncating the sum (5.2) to

$$P_{R_k}(\boldsymbol{X}) = \sum_{i=1}^{\min\{r,k\}} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^t.$$

When $r > k$ and $\sigma_{k+1} = \sigma_k$, the projection operator is multi-valued.

The MM principle allows one to restore the symmetry lost in the missing entries [34]. Suppose $\boldsymbol{X}_n$ is the current approximation to $\boldsymbol{X}$. One simply replaces a missing entry y_{ij} of

$\boldsymbol{Y}$ for $(i,j) \notin \Delta$ by the corresponding entry x_{nij} of $\boldsymbol{X}_n$ and adds the term $\frac{1}{2}(x_{nij} - x_{ij})^2$ to the least squares criterion

$$f(\boldsymbol{X}) = \frac{1}{2} \sum_{(i,j)\in\Delta} (y_{ij} - x_{ij})^2.$$

Since the added terms majorize 0, they create a legitimate surrogate function. One can rephrase the surrogate by defining the orthogonal complement operator $P_\Delta^\perp(\boldsymbol{Y})$ via the equation $P_\Delta^\perp(\boldsymbol{Y}) + P_\Delta(\boldsymbol{Y}) = \boldsymbol{Y}$. The matrix $\boldsymbol{Z}_n = P_\Delta(\boldsymbol{Y}) + P_\Delta^\perp(\boldsymbol{X}_n)$ temporarily completes $\boldsymbol{Y}$ and yields the surrogate function $\frac{1}{2}\|\boldsymbol{Z}_n - \boldsymbol{X}\|_F^2$. In implementing a slightly modified version of the proximal distance algorithm, one must solve for the minimum of the Moreau function

$$\frac{1}{2}\|\boldsymbol{Z}_n - \boldsymbol{X}\|_F^2 + \frac{w_n}{2}\|\boldsymbol{X} - P_{R_k}(\boldsymbol{X}_n)\|_F^2.$$

The stationarity condition

$$\boldsymbol{0} = \boldsymbol{X} - \boldsymbol{Z}_n + w_n[\boldsymbol{X} - P_{R_k}(\boldsymbol{X}_n)]$$

yields the trivial solution

$$\boldsymbol{X}_{n+1} = \frac{1}{1+w_n}\boldsymbol{Z}_n + \frac{w_n}{1+w_n}P_{R_k}(\boldsymbol{X}_n).$$

Again this is guaranteed to decrease the objective function

$$F_\rho(\boldsymbol{X}) = \frac{1}{2} \sum_{(i,j)\in\Delta} (y_{ij} - x_{ij})^2 + \frac{\rho}{2} \operatorname{dist}_\epsilon(\boldsymbol{X}, R_k)$$

for the choice $w_n = \rho/\operatorname{dist}_\epsilon(\boldsymbol{X}_n, R_k)$.

p	q	α	rank	L_1	L_1/L_2	T_1	T_1/T_2
200	250	0.05	20	1598	0.251	4.66	7
800	1000	0.20	80	571949	0.253	131.02	18.1
1000	1250	0.25	100	1112604	0.24	222.2	15.1
1200	1500	0.15	40	793126	0.361	161.51	3.6
1200	1500	0.30	120	1569105	0.235	367.78	12.3
1400	1750	0.35	140	1642661	0.236	561.76	9
1800	2250	0.45	180	2955533	0.171	1176.22	10.1
2000	2500	0.10	20	822673	0.50	307.89	1.9
2000	2500	0.50	200	1087404	0.192	2342.32	2
5000	5000	0.05	30	7647707	0.664	1827.16	2

Table 5.6. Comparison of the MM proximal distance algorithm to SoftImpute. Abbreviations: p is the number of rows, q is the number of columns, α is the sparsity level, L_1 is the optimal loss under MM, L_2 is the optimal loss under SoftImpute, T_1 is the total computation time (in seconds) for MM, and T_2 is the total computation time for SoftImpute.

In the spirit of Example 5.4, let us derive a local Lipschitz constant based on the value $f(\mathbf{0}) = \frac{1}{2}\sum_{(i,j)\in\Delta} y_{ij}^2$. The inequality

$$\frac{1}{2}\sum_{(i,j)\in\Delta} y_{ij}^2 < \frac{1}{2}\sum_{(i,j)\in\Delta}(y_{ij}-x_{ij})^2 = \frac{1}{2}\sum_{(i,j)\in\Delta}(y_{ij}^2 - 2y_{ij}x_{ij} + x_{ij}^2)$$

is equivalent to the inequality

$$2\sum_{(i,j)\in\Delta} y_{ij}x_{ij} < \sum_{(i,j)\in\Delta} x_{ij}^2.$$

In view of the Cauchy-Schwarz inequality

$$\sum_{(i,j)\in\Delta} y_{ij}x_{ij} \le \sqrt{\sum_{(i,j)\in\Delta} y_{ij}^2}\sqrt{\sum_{(i,j)\in\Delta} x_{ij}^2},$$

no solution $\boldsymbol{x}$ of the constrained problem can satisfy

$$\sqrt{\sum_{(i,j)\in\Delta} x_{ij}^2} > 2\sqrt{\sum_{(i,j)\in\Delta} y_{ij}^2}.$$

When the opposite inequality holds,

$$\|\nabla f(\boldsymbol{x})\|_F = \sqrt{\sum_{(i,j)\in\Delta}(x_{ij}-y_{ij})^2} \le \sqrt{\sum_{(i,j)\in\Delta} x_{ij}^2} + \sqrt{\sum_{(i,j)\in\Delta} y_{ij}^2} \le 3\sqrt{\sum_{(i,j)\in\Delta} y_{ij}^2}.$$

Again this tends to be a conservative estimate of the required local bound on ρ. Table 5.6 compares the performance of the MM proximal distance algorithm and a MATLAB implementation of SoftImpute [34]. Although the proximal distance algorithm is noticeably slower, it substantially lowers the optimal loss and improves in relative speed as problem dimensions grow.

Example 5.7 (Sparse Inverse Covariance Estimation). The graphical lasso has applications in estimating sparse inverse covariance matrices [19]. In this context, one minimizes the convex criterion

$$-\ln\det\boldsymbol{\Theta} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Theta}) + \rho\|\boldsymbol{\Theta}\|_1,$$

where $\boldsymbol{\Theta}^{-1}$ is a $p\times p$ theoretical covariance matrix, $\boldsymbol{S}$ is a corresponding sample covariance matrix, and the graphical lasso penalty $\|\boldsymbol{\Theta}\|_1$ equals the sum of the absolute values of the off-diagonal entries of $\boldsymbol{\Theta}$. The solution exhibits both sparsity and shrinkage. One can avoid shrinkage by minimizing

$$f(\boldsymbol{\Theta}) = -\ln\det\boldsymbol{\Theta} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Theta})$$

subject to $\boldsymbol{\Theta}$ having at most $2k$ nonzero off-diagonal entries. Let T_k^p be the closed set of $p\times p$ symmetric matrices possessing this property. Projection of a symmetric matrix $\boldsymbol{M}$ onto T_k^p can be achieved by arranging the above-diagonal entries of $\boldsymbol{M}$ in decreasing absolute value

and replacing all but the first k of these entries by 0. The below-diagonal entries are treated similarly.

The proximal distance algorithm for minimizing $f(\boldsymbol{\Theta})$ subject to the set constraints operates through the convex surrogate

$$g(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}_n) = f(\boldsymbol{\Theta}) + \frac{w_n}{2}\|\boldsymbol{\Theta} - P_{T_k^p}(\boldsymbol{\Theta}_n)\|_F^2$$
$$w_n = \frac{\rho}{\sqrt{\|\boldsymbol{\Theta}_n - P_{T_k^p}(\boldsymbol{\Theta}_n)\|_F^2 + \epsilon}}.$$

A stationary point minimizes the surrogate and satisfies

$$\mathbf{0} = -\boldsymbol{\Theta}^{-1} + w_n\boldsymbol{\Theta} + \boldsymbol{S} - w_n P_{T_k^p}(\boldsymbol{\Theta}_n). \tag{5.3}$$

If the constant matrix $\boldsymbol{S} - w_n P_{T_k^p}(\boldsymbol{\Theta}_n)$ has spectral decomposition $\boldsymbol{U}_n\boldsymbol{D}_n\boldsymbol{U}_n^*$, then multiplying equation (5.3) on the left by $\boldsymbol{U}_n^*$ and on the right by $\boldsymbol{U}_n$ gives

$$\mathbf{0} = -\boldsymbol{U}_n^*\boldsymbol{\Theta}^{-1}\boldsymbol{U}_n + w_n\boldsymbol{U}_n^*\boldsymbol{\Theta}\boldsymbol{U}_n + \boldsymbol{D}_n.$$

This suggests that we take $\boldsymbol{E} = \boldsymbol{U}_n^*\boldsymbol{\Theta}\boldsymbol{U}_n$ to be diagonal and require its diagonal entries e_i to satisfy

$$0 = -\frac{1}{e_i} + w_n e_i + d_{ni}.$$

Multiplying this identity by e_i and solving for the positive root of the resulting quadratic yields

$$e_i = \frac{-d_{ni} + \sqrt{d_{ni}^2 + 4w_n}}{2w_n}.$$

Given the solution matrix $\boldsymbol{E}_{n+1}$, we reconstruct $\boldsymbol{\Theta}_{n+1}$ as $\boldsymbol{U}_n\boldsymbol{E}_{n+1}\boldsymbol{U}_n^*$.

Finding a local Lipschitz constant is more challenging in this example. Because the identity matrix is feasible, the minimum cannot exceed

$$-\ln\det\boldsymbol{I} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{I}) = \operatorname{tr}(\boldsymbol{S}) = \sum_{i=1}^{p}\omega_i,$$

where $\boldsymbol{S}$ is assumed positive definite with eigenvalues ω_i ordered from largest to smallest. If the candidate matrix $\boldsymbol{\Theta}$ is positive definite with ordered eigenvalues λ_i, then the von Neumann-Fan inequality [6] implies

$$f(\boldsymbol{\Theta}) \geq -\sum_{i=1}^{p}\ln\lambda_i + \sum_{i=1}^{p}\lambda_i\omega_{p-i+1}. \tag{5.4}$$

To show that $f(\boldsymbol{\Theta}) > f(\boldsymbol{I})$ whenever any λ_i falls outside a designated interval, note that the contribution $-\ln\lambda_j + \lambda_j\omega_{p-j+1}$ to the right side of inequality (5.4) is bounded below by $\ln\omega_{p-j+1} + 1$ when $\lambda_j = \omega_{p-j+1}^{-1}$. Hence, $f(\boldsymbol{\Theta}) > f(\boldsymbol{I})$ whenever

$$-\ln\lambda_i + \lambda_i\omega_{p-i+1} > \sum_{i=1}^{p}\omega_i - \sum_{j\neq i}(\ln\omega_{p-j+1} + 1). \tag{5.5}$$

Given the strict convexity of the function $-\ln \lambda_i + \lambda_i \omega_{p-i+1}$, equality holds in inequality (5.5) at exactly two points $\lambda_{i\min} > 0$ and $\lambda_{i\max} > \lambda_{i\min}$. These roots can be readily extracted by bisection or Newton's method. The strict inequality $f(\boldsymbol{\Theta}) > f(\boldsymbol{I})$ holds when any λ_i falls to the left of $\lambda_{i\min}$ or to the right of $\lambda_{i\max}$. Within the intersection of the intervals $[\lambda_{i\max}, \lambda_{i\min}]$, the gradient of $f(\boldsymbol{\Theta})$ satisfies

$$\|\nabla f(\boldsymbol{\Theta})\|_F \leq \|\boldsymbol{\Theta}^{-1}\|_F + \|\boldsymbol{S}\|_F \leq \sqrt{\sum_{i=1}^{p} \lambda_i^{-2}} + \|\boldsymbol{S}\|_F \leq \sqrt{\sum_{i=1}^{p} \lambda_{i\min}^{-2}} + \|\boldsymbol{S}\|_F.$$

This bound serves as a local Lipschitz constant near the optimal point.

p	k_t	k_1	k_2	ρ	L_1	$L_2 - L_1$	T_1	T_1/T_2
8	18	14.0	14.0	0.00186	−12.35	0.01	0.022	43.458
16	42	30.5	28.7	0.00305	−25.17	0.08	0.026	43.732
32	90	53.5	49.9	0.00330	−50.75	0.17	0.054	31.639
64	186	97.8	89.3	0.00445	−98.72	0.53	0.234	28.542
128	378	191.6	169.9	0.00507	−196.09	1.14	1.060	18.693
256	762	345.0	304.2	0.00662	−369.62	2.55	4.253	9.559
512	1530	636.4	566.8	0.00983	−641.89	6.72	19.324	5.679

Table 5.7. Numerical results for precision matrix estimation. Abbreviations: p for matrix dimension, k_t for the number of nonzero entries in the true model, k_1 for the number of true nonzero entries recovered by the MM algorithm, k_2 for the number of true nonzero entries recovered by `glasso`, ρ the average tuning constant for `glasso` for a given k_t, L_1 the average loss from the MM algorithm, $L_1 - L_2$ the difference between L_1 and the average loss from `glasso`, T_1 the average compute time in seconds for the MM algorithm, and T_1/T_2 the ratio of T_1 to the average compute time for `glasso`.

Table 5.7 compares the performance of the MM algorithm to that of the R `glasso` package [19]. The sample precision matrix $\boldsymbol{S}^{-1} = \boldsymbol{L}\boldsymbol{L}^* + \delta \boldsymbol{M}\boldsymbol{M}^*$ was generated by filling the diagonal and first three subdiagonals of the banded lower triangular matrix $\boldsymbol{L}$ with standard normal deviates. Filling $\boldsymbol{M}$ with standard normal deviates and choosing $\delta = 0.01$ imposed a small amount of noise obscuring the band nature of $\boldsymbol{L}\boldsymbol{L}^*$. All table statistics represent averages over 10 runs started at $\boldsymbol{\Theta} = \boldsymbol{S}^{-1}$ with k equal to the true number of nonzero entries in $\boldsymbol{L}\boldsymbol{L}^*$. The MM algorithm performs better in minimizing average loss and recovering nonzero entries.

6. Discussion

The MM principle offers a unique and potent perspective on high-dimensional optimization. The current survey emphasizes proximal distance algorithms and their applications in nonlinear programming. Our construction of this new class of algorithms relies on the exact penalty method of Clarke [10] and majorization of a smooth approximation to the Euclidean distance to the constraint set. Well-studied proximal maps and Euclidean projections constitute the key ingredients of seven realistic examples. These examples illustrate the versatility of the method in handling nonconvex constraints, its improvement as problem dimension

increases, and the pitfalls in sending the tuning constants ρ and ϵ too quickly to their limits. Despite the latter concern, we are sufficiently encouraged to pursue this research further, particularly in statistical applications where model fitting and selection are compromised by aggressive penalization.

Acknowledgments. Kenneth Lange was supported by NIH grants from the National Human Genome Research Institute (HG006139) and the National Institute of General Medical Sciences (GM053275). Kevin L. Keys was supported by a National Science Foundation Graduate Research Fellowship under Grant Number DGE-0707424.

References

[1] Baum LE, *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*, Inequalities **3** (1972), 1–8.

[2] Bauschke HH, Borwein JM, and Li W, *Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization*, Math Programming, Series A **86** (1999), 135–160.

[3] Bauschke HH and Combettes PL, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2011.

[4] Borg I and Groenen PJF, *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York, 2007.

[5] Boyd S and Vandenberghe L, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[6] Borwein JM and Lewis AS, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer, New York, 2000.

[7] Bregman LM, *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, USSR Computational Math and Mathematical Physics **7** (1967), 200–217.

[8] Candès EJ and Recht B, *Exact matrix completion via convex optimization*, Foundations Computational Math. **9** (2009), 717–772.

[9] Chi E, Zhou H, and Lange K, *Distance majorization and its applications*, Math. Programming Series A (2013) (in press).

[10] Clarke FH, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, 1983.

[11] Demyanov VF, *Nonsmooth optimization*, in Nonlinear Optimization (editors Di Pillo G, Schoen F), Springer, New York, 2010.

[12] Demyanov VF, Di Pillo G, and Facchinei F, *Exact penalization via Dini and Hadamard conditional derivatives*, Optimization Methods and Software **9** (1998), 19–36.

[13] de Leeuw J, *Applications of convex analysis to multidimensional scaling*, Recent Developments in Statistics, edited by Barra JR, Brodeau F, Romier G, van Cutsem B, North Holland Publishing Company, 1977, pp. 133–146.

[14] _____, *Multivariate analysis with optimal scaling* (1990), Progress in Multivariate Analysis, edited by Das Gupta S, Sethuraman J, Indian Statistical Institute.

[15] Dempster AP, Laird NM, and Rubin DB, *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, J Roy Stat Soc B **39** (1977), 1–38.

[16] Deutsch F, *Best Approximation in Inner Product Spaces*, Springer, New York, 2001.

[17] Dinkelbach W, *On nonlinear fractional programming*, Management Science **13** (1967), 492–498.

[18] Dykstra RL, *An algorithm for restricted least squares estimation*, JASA **78** (1983), 837–842.

[19] Friedman J, Hastie T, and Tibshirani R, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics **9** (2008), 432–441.

[20] Golub GH and Van Loan CF, *Matrix Computations*, 3rd ed. Johns Hopkins University Press, Baltimore, MD, 1996.

[21] Grant MC and Boyd S, *CVX: Matlab software for disciplined convex programming*, version 2.0 beta, 2013.

[22] Hartley HO, *Maximum likelihood estimation from incomplete data*, Biometrics **14** (1958), 174–194.

[23] Heiser WJ, *Convergent computing by iterative majorization: theory and applications in multidimensional data analysis*, Recent Advances in Descriptive Multivariate Analysis, edited by Krzanowski WJ, Oxford University Press, 1995, pp. 157–189.

[24] Hoffman AJ, *On approximate solutions of systems of linear inequalities*, J Res Nat Bur Stand **49** (1952), 263–265.

[25] Hunter DR, Lange K, *A tutorial on MM algorithms*, Amer Statistician **58** (2004), 30–37.

[26] Kiers H, *Majorization as a tool for optimizing a class of matrix functions*, Psychometrika **55** (1990), 417–428.

[27] Lange K, *An adaptive barrier method for convex programming*, Methods Applications Analysis **1** (1994), 392–402.

[28] Lange K, Hunter D, and Yang I, *Optimization transfer using surrogate objective functions (with discussion)*, J Computational Graphical Stat. **9** (2000), 1–59.

[29] Lange K, *Numerical Analysis for Statisticians*, 2nd ed., Springer, 2010.

[30] _____, *Optimization*, 2nd ed., Springer, 2013.

[31] Löfberg J, *YALMIP : A Toolbox for Modeling and Optimization in MATLAB*, Proceedings of the 2004 CACSD Conference, Taipei, Taiwan, 2004.

[32] McKendrick AG, *Applications of mathematics to medical problems*, Proc. Edinburgh Math. Soc. **44** (1926), 1–34.

[33] McLachlan GJ and Krishnan T, *The EM Algorithm and Extensions,* 2nd ed., Wiley, Hoboken, NJ, 2008.

[34] Mazumder R, Hastie T, and Tibshirani R, *Spectral regularization algorithms for learning large incomplete matrices*, J Machine Learning Res. **11** (2010), 2287–2322.

[35] Nesterov Y and Nemirovski A, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[36] Ortega JM and Rheinboldt WC, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic, 1970, pp. 253–255.

[37] Parikh N and Boyd S, *Proximal algorithms*, Foundations Trends Optimization **1** (2013), 123–231.

[38] Smith CAB, *Counting methods in genetical statistics*, Ann Hum Genet **21** (1957), 254–276.

[39] Sundberg R, *An iterative method for solution of the likelihood equations for incomplete data from exponential families*, Communications Stat B **5** (1976), 55–64.

[40] Weiszfeld, E, (1937) *On the point for which the sum of the distances to n given points is minimum*, Ann Oper Research **167**:7–41. Translated from the French original in Tohoku Math J 43:335–386 (1937) and annotated by Frank Plastria.

[41] Yates F, *The analysis of multiple classifications with unequal numbers in different classes*, J Amer Stat. Assoc. **29** (1934), 51–66.

University of California, Los Angeles
E-mail: klange@ucla.edu

University of California, Los Angeles
E-mail: klkeys@ucla.edu

Heat flows, geometric and functional inequalities

Michel Ledoux

Abstract. Heat flow and semigroup interpolations have developed over the years as a major tool for proving geometric and functional inequalities. Main illustrations presented here range over logarithmic Sobolev inequalities, heat kernel bounds, isoperimetric-type comparison theorems, Brascamp-Lieb inequalities and noise stability. Transportation cost inequalities from optimal mass transport are also part of the picture as consequences of new Harnack-type inequalities. The geometric analysis involves Ricci curvature lower bounds via, as a cornerstone, equivalent gradient bounds on the diffusion semigroups. Most of the results presented here are joint with D. Bakry.

Mathematics Subject Classification (2010). Primary 35K05, 39B62, 47D07, 53C21; Secondary 60J60, 58J65.

Keywords. Heat flow, Markov diffusion semigroup, geometric and functional inequality, curvature bound, gradient bound, optimal transport, noise stability.

1. Introduction

The last decades have seen important developments of heat flow methods towards a variety of geometric and functional inequalities. Heat flow or semigroup interpolation is a classical analytic tool, going back at least as far as the so-called Duhamel formula, which has been widely used in a number of settings. The modern era, starting in the eighties, emphasized dynamical proofs of Euclidean and Riemannian functional and geometric inequalities under curvature bounds, as put forward in the early contribution [7] by D. Bakry and M. Émery (see also [6]) in the context of hypercontractivity and logarithmic Sobolev inequalities for diffusion operators. The picture encircles today inequalities relevant to heat kernel and gradient bounds, geometric comparison theorems, Sobolev embeddings, convergence to equilibrium, optimal transport, isoperimetry and measure concentration (as illustrated e.g. in [9]). This text surveys some of these achievements with a particular focus on Sobolev-type, isoperimetric and multilinear inequalities, and noise stability.

Section 2 is a first illustration of the power of heat flow monotonicity towards logarithmic Sobolev inequalities, including in the same picture the classical parabolic Li-Yau inequality. Section 3 presents more refined isoperimetric-type inequalities, leading to comparison of the isoperimetric profile of (infinite-dimensional) curved models with the Gaussian profile. Harnack inequalities drawn from heat flow provide links with optimal mass transport and transportation cost inequalities illustrated in Section 4. The classical Brascamp-Lieb inequalities for multilinear integrals of products of functions form another important family of functional and geometric inequalities. While classically analyzed as isoperimetric inequalities by rear-

Proceedings of the International Congress of Mathematicians, Seoul, 2014

rangement methods, recent developments using semigroup interpolation shed new light on their structure and extremizers. The last Section 6 presents some recent progress connecting even further Brascamp-Lieb and isoperimetric inequalities via (Gaussian) noise stability.

One natural framework of investigation is Euclidean space $\mathbb{R}^n$ or a (weighted) Riemannian manifold in which case (Ricci) curvature lower-bounds enter into the picture. Based upon the early achievement [7] (see [6]), the more general setting of Markov Triples (E, μ, Γ) allows us to develop semigroup interpolation in a wide context, concentrating on the basic algebraic Γ-calculus underlying many of the heat flow arguments. The iterated carré du champ operator Γ_2 provides here the natural functional interpretation of the geometric Bochner formula and of curvature-dimension conditions. The recent book [9] gives an overview of semigroup methods in the context of Markov Triples and their applications to functional and geometric inequalities. Most of the results emphasized here are developed in this monograph [9] written jointly with D. Bakry and I. Gentil, to which we refer for further motivation and illustrations.

2. Logarithmic Sobolev and parabolic Li-Yau inequalities

The celebrated logarithmic Sobolev inequality of L. Gross [47], comparing entropy and Fisher information, is one prototypical example of functional inequality which may be investigated by heat flow methods. Let $d\gamma(x) = (2\pi)^{-n/2} e^{-|x|^2/2} dx$ be the standard Gaussian measure on the Borel sets of $\mathbb{R}^n$.

Theorem 2.1 (Gross' logarithmic Sobolev inequality). *For any smooth positive function $f : \mathbb{R}^n \to \mathbb{R}$ such that $\int_{\mathbb{R}^n} f d\gamma = 1$,*

$$\int_{\mathbb{R}^n} f \log f d\gamma \leq \frac{1}{2} \int_{\mathbb{R}^n} \frac{|\nabla f|^2}{f} d\gamma.$$

Logarithmic Sobolev inequalities are infinite-dimensional counterparts of the classical Sobolev inequalities, and characterize smoothing properties in the form of hypercontractivity. They prove central in a variety of contexts, including entropic convergence to equilibrium of solutions of evolutionary partial differential equations and of Markov chains and models from statistical mechanics, infinite-dimensional Gaussian analysis and measure concentration (cf. e.g. [6, 9, 52, 76] and the references therein).

While there are numerous different proofs of Gross' logarithmic Sobolev inequality, the perhaps simplest one, put forward by D. Bakry and M. Émery [7] in the mid-eighties, uses semigroup interpolation. Indeed, consider the basic (convolution) heat semigroup $(P_t)_{t \geq 0}$ on $\mathbb{R}^n$ given on suitable functions $f : \mathbb{R}^n \to \mathbb{R}$ by

$$P_t f(x) = \frac{1}{(4\pi t)^{n/2}} \int_{\mathbb{R}^n} f(y)\, e^{-|x-y|^2/4t} dy, \quad t > 0,\ x \in \mathbb{R}^n.$$

Given the initial condition f, $u = u(t, x) = P_t f(x)$ solves the heat equation $\partial_t u = \Delta u$ with thus $u(0, x) = f(x)$.

Towards the logarithmic Sobolev inequality of Theorem 2.1, consider the entropy of a positive smooth function f on $\mathbb{R}^n$ along the semigroup $(P_t)_{t \geq 0}$ given by, at any $t > 0$ and any point x (omitted below),

$$P_t(f \log f) - P_t f \log P_t f.$$

The heat flow interpolation then amounts to

$$P_t(f\log f) - P_t f \log P_t f \;=\; \int_0^t \frac{d}{ds}\, P_s(P_{t-s} f \log P_{t-s} f)ds.$$

By the heat equation and the chain rule formula, both in time and space, for $s < t$,

$$\frac{d}{ds} P_s(P_{t-s} f \log P_{t-s} f) \;=\; P_s\left(\frac{|\nabla P_{t-s} f|^2}{P_{t-s} f}\right) \;=\; \phi(s).$$

As gradient and semigroup commute $\nabla P_u f = P_u(\nabla f)$, for every $u > 0$, by the Jensen and Cauchy-Schwarz inequalities (along the Markov operator P_u),

$$|\nabla P_u f|^2 \;\le\; \big[P_u\big(|\nabla f|\big)\big]^2 \le P_u\left(\frac{|\nabla f|^2}{f}\right) P_u f. \tag{2.1}$$

With $u = t - s$, it follows that

$$\phi(s) \;\le\; P_s\, P_{t-s}\left(\frac{|\nabla f|^2}{f}\right) \;=\; \phi(t)$$

so that

$$P_t(f\log f) - P_t f \log P_t f \;=\; \int_0^t \phi(s)ds \;\le\; t\phi(t) \;=\; t\, P_t\left(\frac{|\nabla f|^2}{f}\right). \tag{2.2}$$

When $t = \frac{1}{2}$, this heat kernel (that is, along the distribution of P_t) inequality is precisely, by homogeneity, the Gross logarithmic Sobolev inequality of Theorem 2.1.

It is a significant observation that the preceding argument may be reversed. Indeed, with $u = s$ and f replaced by $P_{t-s} f$, it holds similarly that $\phi(s) \ge \phi(0)$ so that

$$P_t(f\log f) - P_t f \log P_t f \;\ge\; t\,\phi(0) = t\,\frac{|\nabla P_t f|^2}{P_t f}\,. \tag{2.3}$$

This reverse inequality is a relevant property leading to gradient bounds (see below and [9]).

The preceding analysis actually shows that the map

$$s \in [0,t] \;\mapsto\; \phi(s) \;=\; P_s\left(\frac{|\nabla P_{t-s} f|^2}{P_{t-s} f}\right) \;=\; P_s\big(P_{t-s} f |\nabla \log P_{t-s} f|^2\big)$$

is increasing. Following [7], an alternative approach to this fact is of course to take derivative (that is, the second derivative of entropy) yielding

$$\phi'(s) = 2\, P_s\big(P_{t-s} f\, \Gamma_2\,(\log P_{t-s} f)\big)$$

where the Γ_2 operator is given, on any smooth function h on $\mathbb{R}^n$, by

$$\Gamma_2(h) \;=\; \frac{1}{2}\,\Delta\big(|\nabla h|^2\big) - \nabla h \cdot \nabla(\Delta h) \;=\; \big|\mathrm{Hess}(h)\big|^2 \;\ge\; 0.$$

Hence $\phi'(s) \ge 0$ and ϕ is increasing.

The same formalism also works in an n-dimensional Riemannian manifold (M, g) along the heat semigroup $(P_t)_{t\geq 0}$ with Laplace-Beltrami operator Δ as infinitesimal generator. In this case, by the classical Bochner identity, the Γ_2 operator takes the form

$$\Gamma_2(h) \,=\, \mathrm{Ric}_g(\nabla h, \nabla h) + \big|\mathrm{Hess}(h)\big|^2$$

where Ric_g denotes the Ricci tensor of the metric g. Whenever (M, g) has non-negative Ricci curvature, we have similarly that $\phi' \geq 0$ yielding the preceding heat kernel inequalities (2.2) and (2.3) in this more general context.

Actually, under $\mathrm{Ric}_g \geq 0$, by the trace inequality,

$$\Gamma_2(h) \,=\, \mathrm{Ric}_g(\nabla h, \nabla h) + \big|\mathrm{Hess}(h)\big|^2 \geq \big|\mathrm{Hess}(h)\big|^2 \geq \frac{1}{n}\,(\Delta h)^2. \tag{2.4}$$

Thus

$$\phi'(s) \,\geq\, \frac{2}{n}\, P_s\big(P_{t-s} f\, [\Delta \log P_{t-s} f]^2\big)$$

retaining dimensional information. A somewhat more involved integration then yields a strengthened dimensional logarithmic Sobolev inequality

$$P_t(f \log f) - P_t f \log P_t f \,\leq\, t\, \Delta P_t f + \frac{n}{2}\, P_t f \log\left(1 - \frac{2t}{n}\, \frac{P_t(f \Delta \log f)}{P_t f}\right)$$

(for f a positive smooth function on M). Of more interest is actually the reverse form, analogue of (2.3),

$$P_t(f \log f) - P_t f \log P_t f \,\geq\, t\, \Delta P_t f - \frac{n}{2}\, P_t f \log\Big(1 + \frac{2t}{n}\, \Delta(\log P_t f)\Big).$$

The latter entails implicitly (and explicitly from the proof) that $1 + \frac{2t}{n}\, \Delta(\log P_t f) > 0$, or equivalently the famous Li-Yau parabolic inequality [55], initially established by the maximum principle and embedded here in a heat flow argument [11].

Theorem 2.2 (Li-Yau parabolic inequality). *For any (smooth) positive function f on a Riemannian manifold (M, g) with non-negative Ricci curvature,*

$$\frac{|\nabla P_t f|^2}{(P_t f)^2} - \frac{\Delta P_t f}{P_t f} \,\leq\, \frac{n}{2t}\,.$$

The Li-Yau parabolic inequality has numerous important applications (cf. [39, 55]), in particular to Harnack inequalities of the type

$$P_t f(x) \,\leq\, P_{t+s} f(y) \Big(\frac{t+s}{t}\Big)^{n/2} e^{d(x,y)^2/4s} \tag{2.5}$$

for $f : M \to \mathbb{R}$ positive and $t, s > 0$, where $d(x, y)$ is the Riemannian distance between $x, y \in M$.

Parallelisms with the Li-Yau gradient estimates and the Perelman F and W entropy functionals (see [65]) are mentioned in the recent contribution [36] of T. Colding where further monotonicity formulas for Ricci curvature with accompanying rigidity theorems are developed (see also [37]).

The preceding heat flow monotonicity principle yielding both logarithmic Sobolev and Li-Yau inequalities may be developed similarly in the extended setting of a weighted n-dimensional Riemannian manifold (M, g) with a weighted measure $d\mu = e^{-V} dx$, where V is a smooth potential on M, invariant and symmetric with respect to the operator

$$\mathrm{L} = \Delta - \nabla V \cdot \nabla$$

for which the Ricci tensor is extended into the so-called Bakry-Émery tensor $\mathrm{Ric}_g + \mathrm{Hess}(V)$. On the basis of Bochner's identity and (2.4), curvature-dimension $CD(K, N)$ conditions

$$\Gamma_2(h) \;=\; \big[\mathrm{Ric}_g + \mathrm{Hess}(V)\big](\nabla h, \nabla h) + \big|\mathrm{Hess}(h)\big|^2 \;\geq\; K|\nabla h|^2 + \frac{1}{N}\,(\mathrm{L}h)^2 \qquad (2.6)$$

for every smooth h on (M, g), where $K \in \mathbb{R}$ and $N \geq n$ (not necessarily the topological dimension), encode Ricci curvature lower bounds and dimension. The condition (2.6) is inspired by Lichnerowicz' eigenvalue lower bound [9, 45, 56]. Similar functional and heat kernel inequalities are then achieved under $CD(0, N)$ and also $CD(K, N)$.

The results furthermore extend to the general setting of abstract Markov diffusion operators leading to the concept of Markov Triple [6, 9]. A Markov (diffusion) Triple (E, μ, Γ) consists of a state space E equipped with a diffusion semigroup $(P_t)_{t\geq 0}$ with infinitesimal generator L, carré du champ operator Γ and invariant and reversible σ-finite measure μ. The generator L and the carré du champ operator Γ are intrinsically related by the formula

$$\Gamma(f, g) = \frac{1}{2}\big[\mathrm{L}(fg) - f\,\mathrm{L}g - g\,\mathrm{L}f\big]$$

for functions f, g belonging to a suitable algebra $\mathcal{A}$ of functions (corresponding to smooth functions in a Riemannian setting). The state space E may be endowed with an intrinsic distance d for which Lipschitz functions f are such that $\Gamma(f)$ is bounded (μ-almost everywhere). In the (weighted) Riemannian context, L is the Laplace operator Δ with drift $-\nabla V \cdot \nabla$ with respect to the weighted measure $d\mu = e^{-V} dx$, $\Gamma(f, f) = |\nabla f|^2$ for smooth functions, and d corresponds to the Riemannian metric.

In the Markov Triple setting, the abstract curvature condition

$$CD(K, N),\; K \in \mathbb{R},\; N \geq 1,$$

mimicking (2.6), takes the form

$$\Gamma_2(h) \;\geq\; K\,\Gamma(h) + \frac{1}{N}\,(\mathrm{L}h)^2, \quad h \in \mathcal{A}, \qquad (2.7)$$

(with the shorthand notation $\Gamma(h) = \Gamma(h, h)$, $\Gamma_2(h) = \Gamma_2(h, h)$) where the Bakry-Émery Γ_2 operator, going back to [7] (see [6, 9]), is defined from Γ by

$$\Gamma_2(f, g) \;=\; \frac{1}{2}\big[\mathrm{L}\big(\Gamma(f, g)\big) - \Gamma(f, \mathrm{L}g) - \Gamma(g, \mathrm{L}f)\big], \quad (f, g) \in \mathcal{A} \times \mathcal{A}.$$

As a major property emphasized by D. Bakry [5, 6, 9], the curvature condition $CD(K, \infty)$ is translated equivalently into gradient bounds, allowing in particular, along (2.1), for the preceding semigroup interpolation arguments and heat kernel inequalities.

Theorem 2.3 (Gradient bound). *The curvature condition $CD(K,\infty)$ holds true if and only if for any $t \geq 0$ and any $f \in \mathcal{A}$,*

$$\sqrt{\Gamma(P_t f)} \leq e^{-Kt} P_t\big(\sqrt{\Gamma(f)}\big).$$

The curvature-dimension condition $CD(K,N)$ leads on the other hand to dimensional gradient bounds of the type [11, 79]

$$\Gamma(P_t f) \leq e^{-2Kt} P_t\big(\Gamma(f)\big) - \frac{1-e^{-2Kt}}{KN} (\mathrm{L}P_t f)^2 \tag{2.8}$$

which are central in the comparison with alternative curvature-dimension conditions from optimal transport (see Section 4).

3. Isoperimetric-type inequalities

More refined isoperimetric statements may be achieved by the preceding semigroup interpolation arguments. One prototypical result in this direction is a comparison theorem between the isoperimetric profile of a curved infinite-dimensional diffusion operator (in the preceding sense) and the Gaussian profile.

Denote by $\mathrm{I} : [0,1] \to \mathbb{R}_+$ the Gaussian isoperimetric function defined by $\mathrm{I} = \varphi \circ \Phi^{-1}$ where

$$\Phi(x) = \int_{-\infty}^{x} e^{-u^2/2} \frac{du}{\sqrt{2\pi}}, \quad x \in \mathbb{R},$$

is the distribution function of a standard normal and $\varphi = \Phi'$ its density. The following theorem ([10]) is presented in the general context of a Markov Triple (E, μ, Γ) (with underlying algebra of smooth functions $\mathcal{A}$), covering in particular the setting of weighted Riemannian manifolds.

Theorem 3.1 (Gaussian isoperimetry for heat kernel measure). *Let (E, μ, Γ) be a Markov Triple satisfying the curvature condition $CD(K,\infty)$ for some $K \in \mathbb{R}$. For every function f in $\mathcal{A}$ with values in $[0,1]$ and every $t \geq 0$,*

$$\mathrm{I}(P_t f) \leq P_t\Big(\sqrt{\mathrm{I}^2(f) + K(t)\,\Gamma(f)}\Big)$$

where $K(t) = \frac{1}{K}(1 - e^{-2Kt})$ $(= 2t$ if $K = 0)$.

For the example of the standard heat semigroup on $\mathbb{R}^n$ with $t = \frac{1}{2}$, Theorem 3.1 yields that for any smooth function $f : \mathbb{R}^n \to [0,1]$,

$$\mathrm{I}\Big(\int_{\mathbb{R}^n} f d\gamma\Big) \leq \int_{\mathbb{R}^n} \sqrt{\mathrm{I}^2(f) + |\nabla f|^2}\, d\gamma. \tag{3.1}$$

This inequality applied to εf as $\varepsilon \to 0$, together with the asymptotics $\mathrm{I}(v) \sim v\sqrt{2\log\frac{1}{v}}$ as $v \to 0$, strengthens the logarithmic Sobolev inequality of Theorem 2.1. A smooth approximation f of the characteristic function of a Borel set A in $\mathbb{R}^n$ ensures that

$$\mathrm{I}\big(\gamma(A)\big) \leq \gamma^+(A) = \liminf_{\varepsilon \to 0} \frac{1}{\varepsilon}\big[\gamma(A_\varepsilon) - \gamma(A)\big] \tag{3.2}$$

where the right-hand side defines the Minkowski content (surface measure) of A (where, for $\varepsilon > 0$, $A_\varepsilon = \{x \in E\,;\, d(x,A) \leq \varepsilon\}$). This inequality (3.2) exactly expresses the isoperimetric problem for the Gaussian measure γ on $\mathbb{R}^n$ for which half-spaces

$$H = \{x \in \mathbb{R}^n\,;\, x \cdot u \leq a\},$$

where u is a unit vector and $a \in \mathbb{R}$, achieve the minimal surface measure at fixed measure. Indeed, if a is chosen such that $\gamma(A) = \Phi(a)$, then $\gamma(A) = \gamma(H)$ and

$$\gamma^+(H) \,=\, \varphi(a) \,=\, \mathrm{I}\big(\Phi(a)\big) \,\leq\, \gamma^+(A).$$

The Gaussian isoperimetric inequality (3.2) goes back to V. Sudakov and B. Tsirel'son [73] and C. Borell [23] relying on the isoperimetric inequality on spheres and a limiting argument. The functional form (3.1) has been put forward by S. Bobkov [21] (see also earlier [42] within Gaussian symmetrization [41]).

The content of Theorem 3.1 is therefore that the isoperimetric profile of the heat kernel measures (of a positively curved diffusion semigroup) is bounded from below, up to a constant, by the isoperimetric profile I of the standard Gaussian measure (in dimension one actually). In particular, if $d\mu = e^{-V} dx$ is a probability measure on $\mathbb{R}^n$ with smooth potential V such that $\mathrm{Hess}(V) \geq K\,\mathrm{Id}$ for some $K > 0$ as symmetric matrices, the curvature condition $CD(K,\infty)$ holds and one may let t tend to ∞ in Theorem 3.1 to see that the isoperimetric profile of μ,

$$\mathrm{I}_\mu(v) \,=\, \inf\big\{\mu^+(A)\,;\, \mu(A) = v\big\}, \quad v \in (0,1),$$

is bounded from below by $\sqrt{K}\,\mathrm{I}$. In this sense, Theorem 3.1 is the infinite-dimensional analogue of the Lévy-Gromov isoperimetric comparison theorem [59] bounding from below the isoperimetric profile of the (normalized) Riemannian measure of an n-dimensional Riemannian manifold with Ricci curvature bounded from below by $n-1$, by the isoperimetric profile of the standard n-sphere. A heat flow proof of this result is yet to be found. For far-reaching geometric generalizations of the Lévy-Gromov theorem, see [58].

In the same spirit as (2.3), reverse forms of the isoperimetric heat kernel inequalities of Theorem 3.1 are also available. Under the curvature condition $CD(K,\infty)$ for some $K \in \mathbb{R}$, for every function f in $\mathcal{A}$ with values in $[0,1]$ and every $t > 0$,

$$\big[\mathrm{I}(P_t f)\big]^2 - \big[P_t\big(\mathrm{I}(f)\big)\big]^2 \,\geq\, \frac{1}{K}\,(e^{2Kt} - 1)\,\Gamma(P_t f).$$

These (sharp) gradient bounds may then be used to prove new isoperimetric-type Harnack inequalities [8].

Theorem 3.2 (Isoperimetric Harnack inequality). *Let (E,μ,Γ) be a Markov Triple satisfying the curvature condition $CD(K,\infty)$ for some $K \in \mathbb{R}$. For every measurable set A in E, every $t \geq 0$ and every $x, y \in E$ such that $d(x,y) > 0$,*

$$P_t(\mathbb{1}_A)(x) \,\leq\, P_t\big(\mathbb{1}_{A_{d_t}}\big)(y)$$

where $d_t = e^{-Kt} d(x,y)$. In particular, when $K = 0$,

$$P_t(\mathbb{1}_A)(x) \,\leq\, P_t\big(\mathbb{1}_{A_{d(x,y)}}\big)(y).$$

Under the curvature condition $CD(K,\infty)$, it is not possible to expect standard (dimensional) Harnack inequalities of the type (2.5). However, the set inequalities of Theorem 3.2 yield infinite-dimensional analogues first obtained by F.-Y. Wang in [77, 78]. For simplicity, the $CD(0,\infty)$ version states the following.

Theorem 3.3 (Wang's Harnack inequality). *In the preceding context, under the curvature condition $CD(0,\infty)$, for every positive (measurable) function f on E, every $t > 0$, every $\alpha > 1$, and every $x, y \in E$,*

$$\big(P_t f(x)\big)^\alpha \,\le\, P_t(f^\alpha)(y)\, e^{\alpha d(x,y)^2/4(\alpha-1)t}. \tag{3.3}$$

In the limit as $\alpha \to \infty$, the latter turns into a log-Harnack inequality

$$P_t(\log f)(x) \,\le\, \log P_t f(y) + \frac{d(x,y)^2}{4t} \tag{3.4}$$

for f positive.

4. Transportation cost inequalities

Heat flow methods have developed simultaneously in the context of transportation cost inequalities which are parts of the main recent achievements in optimal transport (cf. [76]). In particular, they may be used to reach the famous HWI inequality of F. Otto and C. Villani [67] connecting (Boltzmann H-) Entropy, Wasserstein distance (W) and Fisher Information (I).

The HWI inequality covers at the same time logarithmic Sobolev and transportation cost inequalities (in the form of the Talagrand quadratic transportation cost inequality [74]). For simplicity, we deal here with a weighted Riemannian manifold (M, g) with weighted probability measure $d\mu = e^{-V} dx$, and restrict ourselves to the non-negative curvature condition $CD(0,\infty)$. The (quadratic) Wasserstein distance $\mathrm{W}_2(\nu,\mu)$ between two probability measures μ and ν on M is defined by

$$\mathrm{W}_2(\nu,\mu) \,=\, \bigg(\int_{M\times M} d(x,y)^2 d\pi(x,y) \bigg)^{1/2}$$

where the infimum is taken over all couplings π with respective marginals ν and μ (cf. [75, 76]).

Theorem 4.1 (Otto-Villani HWI inequality). *Under the curvature condition $CD(0,\infty)$, for any smooth positive function $f : M \to \mathbb{R}$ with $\int_M f d\mu = 1$,*

$$\int_M f \log f d\mu \,\le\, \mathrm{W}_2(\nu,\mu) \bigg(\int_M \frac{|\nabla f|^2}{f}\, d\mu \bigg)^{1/2}$$

where $d\nu = f d\mu$.

The starting point towards a semigroup proof (first emphasized in [22]) is the log-Harnack inequality (3.4) which may be translated equivalently as

$$P_t(\log f) \le Q_{2t}(\log P_t f) \tag{4.1}$$

where $(Q_s)_{s>0}$ is the Hopf-Lax infimum-convolution semigroup

$$Q_s\varphi(x) \,=\, \inf_{y\in M}\left[\varphi(y) + \frac{d(x,y)^2}{2s}\right], \quad s>0,\ x\in M.$$

This convolution semigroup is closely related to the Wasserstein distance W_2 via the Kantorovich dual description

$$\frac{1}{2}\,\mathrm{W}_2(\nu,\mu)^2 \,=\, \sup\left[\int_M Q_1\varphi\,d\nu - \int_M \varphi\,d\mu\right] \tag{4.2}$$

where the supremum runs over all bounded continuous functions $\varphi : M \to \mathbb{R}$ (cf. [75, 76]).

Given $f>0$ a (smooth bounded) probability density with respect to μ and $d\nu = f d\mu$, simple symmetry and scaling properties on the basis of (4.1) and (4.2) yield that

$$\int_M P_t f \log P_t f d\mu \,\le\, \frac{1}{4t}\,\mathrm{W}_2^2(\nu,\mu). \tag{4.3}$$

The heat flow interpolation scheme illustrated in Section 2 expresses on the other hand that for every $t>0$,

$$\int_M f\log f d\mu \,\le\, \int_M P_t f\log P_t f d\mu + t\int_M \frac{|\nabla f|^2}{f}\,d\mu.$$

Together with (4.3), optimization in $t>0$ yields the HWI inequality. Similar arguments may be developed under $CD(K,\infty)$ for any $K\in\mathbb{R}$ to yield the full formulation of Otto-Villani's HWI inequality (cf. [9, 22]).

The HWI inequality is one important illustration of the description by F. Otto [49, 66] of the heat flow as the gradient flow of entropy, which led to the introduction of curvature lower bounds in metric measure spaces as convexity of entropy along the geodesics of optimal transport by J. Lott, C. Villani [57] and K.-T. Sturm [72] (cf. [76]). Recent major achievements by L. Ambrosio, N. Gigli, G. Savaré [1–3] and M. Erbar, K. Kuwada, K.-T. Sturm [44] establish the equivalence of the curvature and curvature-dimension lower bounds in the sense of the Bakry-Émery Γ_2 operator and of optimal transport in the class of Riemannian Energy (metric) measure spaces with, in particular, the tools of the gradient bounds of Theorem 2.3 and (2.8).

A further by-product of the isoperimetric Harnack Theorem 3.2 in this context is a commutation property between the actions of the heat $(P_t)_{t\ge 0}$ and Hopf-Lax $(Q_s)_{s>0}$ semigroups [8], first emphasized by K. Kuwada [50], at the root of the contraction property in Wasserstein distance along the heat flow [33, 66, 68, 69] (see [75, 76, 78]). The following statement is again restricted, for simplicity, to the non-negative curvature condition.

Theorem 4.2 (Contraction property of the Wasserstein distance). *Under the curvature condition $CD(0,\infty)$, for any $t,s>0$*

$$P_t(Q_s) \,\le\, Q_s(P_t).$$

As a consequence,

$$\mathrm{W}_2(\mu_t,\nu_t) \le \mathrm{W}_2(\mu_0,\nu_0)$$

where $d\mu_t = P_t f d\mu$ and $d\nu_t = P_t g d\mu$, $t\ge 0$, f,g probability densities with respect to μ. Conversely, both properties are equivalent to $CD(0,\infty)$.

5. Brascamp-Lieb inequalities

The Brascamp-Lieb inequalities for multilinear integrals of products of functions in several dimensions were first investigated with rearrangement tools [27, 28]. A later approach, including inverse forms, was developed by F. Barthe via mass transportation [13]. Investigations of E. Carlen, E. Lieb, M. Loss [31] and J. Bennett, A. Carbery, M. Christ, T. Tao [19] promoted heat flow monotonicity as a major tool towards these inequalities and full geometric descriptions of their extremizers.

The basic principle, in a reduced simple instance, is best developed with respect to the so-called Ornstein-Uhlenbeck semigroup $(P_t)_{t\geq 0}$ on $\mathbb{R}^n$ with infinitesimal generator

$$\mathrm{L} = \Delta f - x \cdot \nabla$$

(corresponding therefore to the quadratic potential $V(x) = \frac{1}{2}\,|x|^2$), invariant and symmetric with respect to the standard Gaussian measure γ, and given by the integral representation along suitable functions $f : \mathbb{R}^n \to \mathbb{R}$ by

$$P_t f(x) \,=\, \int_{\mathbb{R}^n} f\big(e^{-t}x + \sqrt{1-e^{-2t}}\, y\big) d\gamma(y), \quad t \geq 0,\ x \in \mathbb{R}^n. \tag{5.1}$$

Let J be a (smooth) real-valued function on some open rectangle $\mathcal{R}$ of $\mathbb{R}^m$. A composition like $J \circ f$ is implicitly meant for functions f with values in $\mathcal{R}$. Let $f = (f_1, \ldots, f_m)$ be a vector of (smooth) functions on $\mathbb{R}^n$ and consider,

$$\psi(t) \,=\, \int_{\mathbb{R}^n} J \circ P_t f \, d\gamma, \quad t \geq 0$$

(where the Ornstein-Uhlenbeck semigroup $(P_t)_{t\geq 0}$ is extended to functions with values in $\mathbb{R}^m$). By the heat equation $\partial P_t f \,=\, \mathrm{L} P_t f$ and integration by parts with respect to the generator L, it holds that

$$\psi'(t) \,=\, -\sum_{k,\ell=1}^m \int_{\mathbb{R}^n} \partial_{k\ell} J \circ P_t f \, \nabla P_t f_k \cdot \nabla P_t f_\ell \, d\gamma.$$

Applied to functions $f_k \,=\, g_k \circ A_k$, $k = 1, \ldots, m$, on $\mathbb{R}^{rn}$, where $g_k : \mathbb{R}^s \to \mathbb{R}$ and A_k is a (constant) $s \times rn$ matrix such that $A_k\, {}^t\!A_k$ is the identity matrix (of $\mathbb{R}^s$), the argument expresses the following conclusion. For $k, \ell = 1, \ldots, m$, set $M_{k\ell} \,=\, A_\ell\, {}^t\!A_k$ (which is an $s \times s$ matrix).

Proposition 5.1. *In the preceding notation, provided the Hessian of J is such that for all vectors v_k in $\mathbb{R}^s$, $k = 1, \ldots, m$,*

$$\sum_{k,\ell=1}^m \partial_{k\ell} J \, M_{k\ell}\, v_k \cdot v_\ell \,\leq\, 0, \tag{5.2}$$

then

$$\int_{\mathbb{R}^{rn}} J(g_1 \circ A_1, \ldots, g_m \circ A_m) d\gamma \,\leq\, J\bigg(\int_{\mathbb{R}^{rn}} g_1 \circ A_1 d\gamma, \ldots, \int_{\mathbb{R}^{rn}} g_m \circ A_m d\gamma\bigg).$$

When $s = 1$, the condition (5.2) amounts to the fact that the Hadamard (point-wise) product $\mathrm{Hess}(J) \circ M$ of the Hessian of J and of the matrix $M = (M_{k\ell})_{1 \le k,\ell \le n}$ is (semi-) negative definite.

This general proposition encircles a number of illustrations of interest. As a first example, take $s = n$ and $r = m = 2$ and let A_1 and A_2 be the $n \times 2n$ matrices $A_1 = (\mathrm{Id}_n; 0_n)$ and $A_2 = (\rho\, \mathrm{Id}_n; \sqrt{1-\rho^2}\, \mathrm{Id}_n)$ where $\rho \in (0,1)$. In this case, the monotonicity condition (5.2) is expressed by the non-positivity of the matrix

$$\begin{pmatrix} \partial_{11} J & \rho\, \partial_{12} J \\ \rho\, \partial_{12} J & \partial_{22} J \end{pmatrix}. \tag{5.3}$$

For instance, if $J(u,v) = u^\alpha v^\beta$, $(u,v) \in (0,\infty)^2$, the condition is fulfilled with

$$\rho^2 \alpha\beta \,\le\, (\alpha - 1)(\beta - 1).$$

For this choice of J, Proposition 5.1 indicates that

$$\begin{aligned} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} g_1^\alpha(x) g_2^\beta\big(\rho x + \sqrt{1-\rho^2}\, y\big) d\gamma(x) d\gamma(y) \\ \le \left(\int_{\mathbb{R}^n} g_1 d\gamma \right)^\alpha \left(\int_{\mathbb{R}^n} g_2 d\gamma \right)^\beta. \end{aligned} \tag{5.4}$$

With $\rho = e^{-t}$, by definition of $P_t g_2$ and duality, the preceding amounts to the famous Nelson hypercontractivity property [64] (for the Ornstein-Uhlenbeck semigroup), equivalent to the logarithmic Sobolev inequality of Theorem 2.1 [9, 47].

Theorem 5.2 (Nelson's hypercontractivity). *Whenever $1 < p < q < \infty$ and $e^{2t} \ge \frac{q-1}{p-1}$, for any measurable function $f : \mathbb{R}^n \to \mathbb{R}$,*

$$\|P_t f\|_q \,\le\, \|f\|_p.$$

This example is actually embedded in the so-called geometric form of the Brascamp-Lieb inequalities emphasized by K. Ball (cf. [12, 15, 19]). For simplicity, consider only the one-dimensional versions $r = s = 1$. Let $A_1, \ldots, A_m$ be unit vectors which decompose the identity in $\mathbb{R}^n$ in the sense that for $0 \le c_k \le 1$, $k = 1, \ldots, m$,

$$\sum_{k=1}^m c_k A_k \otimes A_k \,=\, \mathrm{Id}_n. \tag{5.5}$$

Then, for $J(u_1, \ldots, u_m) = u_1^{c_1} \cdots u_m^{c_m}$ on $(0,\infty)^m$ and $f_k(x) = g_k(A_k \cdot x)$, $g_k : \mathbb{R} \to \mathbb{R}$, $k = 1, \ldots, m$, condition (5.2) of Proposition 5.1 amounts to

$$\sum_{k,\ell=1}^m c_k c_\ell A_k \cdot A_\ell\, v_k v_\ell \,\le\, \sum_{k=1}^m c_k v_k^2 \tag{5.6}$$

for all $v_1, \ldots, v_m \in \mathbb{R}$, which is easily seen to follow from (5.5).

Corollary 5.3 (Geometric Brascamp-Lieb inequality). *Under the decomposition* (5.5)*, for positive measurable functions g_k on $\mathbb{R}$, $k = 1, \ldots, m$,*

$$\int_{\mathbb{R}^n} \prod_{k=1}^m g_k^{c_k}(A_k \cdot x) d\gamma \,\le\, \prod_{k=1}^m \left(\int_{\mathbb{R}} g_k d\gamma \right)^{c_k}.$$

These Brascamp-Lieb inequalities are more classically stated with respect to the Lebesgue measure as

$$\int_{\mathbb{R}^n} \prod_{k=1}^m f_k^{c_k}(A_k \cdot x)dx \ \le \ \prod_{k=1}^m \bigg(\int_{\mathbb{R}} f_k dx \bigg)^{c_k}$$

which is immediately obtained after the change $f_k(x) = g_k(x)e^{-x^2/2}$ (using $\sum_{k=1}^m c_k = n$).

It should be mentioned that inverse Brascamp-Lieb inequalities may also be established along the heat equation as emphasized recently in [34]. C. Borell showed in [25] (see also [14]) that the condition, for non-negative functions f, g, h on $\mathbb{R}^n$ and $\theta \in (0,1)$,

$$h\big(\theta x + (1-\theta)y\big) \ \ge \ f(x)^\theta g(y)^{1-\theta} \tag{5.7}$$

for all $x, y \in \mathbb{R}^n$, is stable under the (standard) heat semigroup $(P_t)_{t\ge 0}$ on $\mathbb{R}^n$ (acting on f, g, h). In the limit as $t \to \infty$, it yields the Prékopa-Leindler theorem indicating that under (5.7),

$$\int_{\mathbb{R}^n} h dx \ \ge \ \bigg(\int_{\mathbb{R}^n} f dx \bigg)^\theta \bigg(\int_{\mathbb{R}^n} g dx \bigg)^{1-\theta}.$$

Specialized to the characteristic functions of sets, this theorem entails the geometric Brunn-Minkowski inequality (in its multiplicative form), and hence the classical isoperimetric inequality in Euclidean space (cf. [46]). C. Borell also provides in [26] the analogous analysis for the Gaussian Brunn-Minkowski and isoperimetric inequalities (as conjectured in [41]).

On the basis of the geometric form Brascamp-Lieb inequalities established by monotonicity along the heat equation, the works [19, 20] of J. Bennett, A. Carbery, M. Christ, T. Tao fully analyze finiteness of constants, structure and existence and uniqueness of centered Gaussian extremals of Euclidean Brascamp-Lieb inequalities (see also [17, 29] for a survey). For applications to the Hausdorff-Young inequality, Euclidean convolution and entropic inequalities, see [18, 34, 38].

One of the motivations of E. Carlen, E. Lieb and M. Loss in [31] was to investigate Brascamp-Lieb and Young inequalities for coordinates on the sphere. Let $\mathbb{S}^{n-1}$ be the standard n-sphere in $\mathbb{R}^n$ and denote by σ the uniform (normalized) measure on it.

Theorem 5.4 (Brascamp-Lieb inequality on the sphere). *Assume that J on $\mathbb{R}^n$, or some open (convex) set in $\mathbb{R}^n$, is separately concave in any two variables. If g_k, $k = 1, \ldots, n$, are, say bounded measurable, functions on $[-1, +1]$, then*

$$\int_{\mathbb{S}^{n-1}} J\big(g_1(x_1), \ldots, g_n(x_n)\big)d\sigma \ \le \ J\bigg(\int_{\mathbb{S}^{n-1}} g_1(x_1)d\sigma, \ldots, \int_{\mathbb{S}^{n-1}} g_n(x_n)d\sigma \bigg).$$

The proof proceeds as the one of Proposition 5.1 along now the heat flow of the Laplace operator $\Delta = \frac{1}{2}\sum_{k,\ell=1}^n (x_k\partial_\ell - x_\ell\partial_k)^2$ on $\mathbb{S}^{n-1}$. The monotonicity condition on J then takes the form

$$\sum_{k,\ell=1}^n \partial_{k\ell} J \, (\delta_{k\ell} - x_k x_\ell) v_k v_\ell \ \le \ 0$$

which is easily seen to be satisfied under concavity of J in any two variables. The case considered in [31] corresponds to $J(u_1, \ldots, u_n) = (u_1 \cdots u_n)^{1/2}$ on $\mathbb{R}^n_+$ leading to

$$\int_{\mathbb{S}^{n-1}} g_1(x_1) \cdots g_n(x_n)d\sigma \ \le \ \bigg(\int_{\mathbb{S}^{n-1}} g_1^2(x_1)d\sigma \bigg)^{1/2} \cdots \bigg(\int_{\mathbb{S}^{n-1}} g_n^2(x_n)d\sigma \bigg)^{1/2}.$$

More general forms under decompositions (5.5) of the identity in Riemannian Lie groups have been studied in [15]. Discrete versions on the symmetric group and multivariate hypergeometric models have been considered analogously [15, 32].

As one further illustration of Proposition 5.1, consider $X = (X_1, \ldots, X_m)$ a centered Gaussian vector on $\mathbb{R}^m$ with covariance matrix $M = A\,{}^tA$ such that $M_{kk} = 1$ for every $k = 1, \ldots, m$. The vector X has the distribution of Ax, $x \in \mathbb{R}^n$, under the standard normal distribution γ on $\mathbb{R}^n$. Applying Proposition 5.1 to the unit vectors ($1 \times n$ matrices) A_k, $k = 1, \ldots, m$, which are the lines of the matrix A and to $f_k(x) = g_k(A_k \cdot x)$, $x \in \mathbb{R}^n$, where $g_k : \mathbb{R} \to \mathbb{R}$, $k = 1, \ldots, m$, with respect to γ, yields that whenever $\mathrm{Hess}(J) \circ M \leq 0$,

$$\mathbb{E}\Big(J\big(g_1(X_1), \ldots, g_m(X_m)\big)\Big) \,\leq\, J\Big(\mathbb{E}\big(g_1(X_1)\big), \ldots, \mathbb{E}\big(g_m(X_m)\big)\Big) \tag{5.8}$$

(under suitable integrability properties on the g_k's). See [34] for the case of Brascamp-Lieb functions J and multidimensional forms.

6. Gaussian noise stability

The study of noise stability (or sensitivity) in Boolean analysis is another field of interest in which links with interpolation along the Ornstein-Uhlenbeck semigroup (for the ideal Gaussian continuous model) were developed. Indeed, as recently demonstrated by E. Mossel and J. Neeman [60, 61], for a suitable choice of the function J, the correlation inequality (5.8) actually entails significant inequalities related to (Gaussian) noise stability and isoperimetry.

Set, for $(u, v) \in [0, 1]^2$ and $\rho \in (0, 1)$,

$$J_\rho^{\mathrm{B}}(u, v) \,=\, \gamma \otimes \gamma\big((x, y) \in \mathbb{R}^2 \,; x \leq \Phi^{-1}(u), \rho x + \sqrt{1 - \rho^2}\, y \leq \Phi^{-1}(v)\big).$$

Equivalently, when $\rho = e^{-t}$, $t > 0$,

$$J_\rho^{\mathrm{B}}(u, v) \,=\, \int_{\mathbb{R}^n} \mathbb{1}_H P_t(\mathbb{1}_K) d\gamma$$

where $(P_t)_{t \geq 0}$ is the Ornstein-Uhlenbeck semigroup (5.1) and H and K are the (parallel) half-spaces

$$H \,=\, \big\{x \in \mathbb{R}^n \,; x_1 \leq \Phi^{-1}(u)\big\}, \quad K \,=\, \big\{x \in \mathbb{R}^n \,; x_1 \leq \Phi^{-1}(v)\big\}.$$

As a main feature, the function J_ρ^{B} is ρ-concave in the sense that the matrix (5.3), which is the Hadamard product of the Hessian of J_ρ^{B} with the covariance matrix of the Gaussian vector $(x, \rho x + \sqrt{1 - \rho^2}\, y)$, is non-positive definite. Proposition 5.1 applied to this function J_ρ^{B} as towards hypercontractivity, or equivalently the multidimensional analogue of (5.8), therefore yields

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} J_\rho^{\mathrm{B}}\Big(f(x), g\big(\rho x + \sqrt{1 - \rho^2}\, y\big)\Big) d\gamma(x) d\gamma(y) \leq\, J_\rho^{\mathrm{B}}\bigg(\int_{\mathbb{R}^n} f\, d\gamma, \int_{\mathbb{R}^n} g\, d\gamma\bigg) \tag{6.1}$$

for every measurable functions $f, g : \mathbb{R}^n \to [0, 1]$. Since $J_\rho^{\mathrm{B}}(0, 0) = J_\rho^{\mathrm{B}}(1, 0) = J_\rho^{\mathrm{B}}(0, 1) = 0$ and $J_\rho^{\mathrm{B}}(1, 1) = 1$, the application of (6.1) to $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$ for Borel sets A, B in

$\mathbb{R}^n$ and the very definition of J_ρ^{B} yield a semigroup proof of Borell's noise stability theorem [60, 61]. This result was initially established via symmetrization with respect to the Gaussian measure introduced by A. Ehrhard [24, 41] (along the rearrangement ideas in Euclidean and spherical spaces [4, 28, 70, 71], see also [16, 30, 54]).

Theorem 6.1 (Borell's noise stability theorem). *For Borel sets $A, B \subset \mathbb{R}^n$, and for every $t \geq 0$,*

$$\int_{\mathbb{R}^n} \mathbb{1}_A P_t(\mathbb{1}_B) d\gamma \leq \int_{\mathbb{R}^n} \mathbb{1}_H P_t(\mathbb{1}_K) d\gamma$$

where $H = \{x_1 \leq a\}$, $K = \{x_1 \leq b\}$ are parallel half-spaces with respectively the same Gaussian measures $\gamma(H) = \gamma(A)$ and $\gamma(K) = \gamma(B)$.

Theorem 6.1 thus expresses that half-spaces are the most noise stable in the sense that they maximize $\int_{\mathbb{R}^n} \mathbb{1}_A P_t(\mathbb{1}_A) d\gamma$ over all Borel sets A in $\mathbb{R}^n$. The new semigroup proof by E. Mossel and J. Neeman [60, 61] was motivated by the equality case and the study of the deficit (see below). It is also connected to the discrete version on the cube $\{-1, +1\}^n$ and the "Majority is Stablest" theorem of [62] in the context of hardness of approximation for Max-Cut in Boolean analysis. While established first via an invariance principle on the basis of Theorem 6.1, a recent purely discrete proof is emphasized in [40].

Classical arguments providing (small time) heat flow descriptions of surface measures may be used to recover the standard Gaussian isoperimetric inequality from Theorem 6.1 [51]. Indeed, it holds true that

$$\gamma^+(A) \geq \limsup_{t \to 0} \sqrt{\frac{\pi}{t}} \left[\gamma(A) - \int_{\mathbb{R}^n} \mathbb{1}_A P_t(\mathbb{1}_A) d\gamma \right]$$

with equality on half-spaces, so that together with Theorem 6.1, $\gamma^+(A) \geq \gamma^+(H)$ if H is a half-space with $\gamma(A) = \gamma(H)$. Besides, a suitable limiting procedure, replacing (f, g) by $(\varepsilon f, \delta g)$ as $\varepsilon, \delta \to 0$, shows that (6.1) contains the hypercontractivity inequality (5.4) (cf. [53]).

Recent investigations study bounds on the deficit in the noise stability Theorem 6.1 and the Gaussian isoperimetric inequality (3.2). While semigroup tools may be used to some extent [60, 61], R. Eldan [43] achieved a complete picture with wider and more refined stochastic calculus tools (improving in particular upon former mass transportation arguments [35]). He showed that, up to a logarithmic factor, given $t > 0$ and a Borel set A, there exists a half-space H with $\gamma(H) = \gamma(A)$ such that

$$\int_{\mathbb{R}^n} \mathbb{1}_H P_t(\mathbb{1}_H) d\gamma - \int_{\mathbb{R}^n} \mathbb{1}_A P_t(\mathbb{1}_A) d\gamma \geq C\big(\gamma(A), t\big)\, \gamma(A \Delta H)^2$$

(and similarly for the isoperimetric deficit), independently of the dimension.

Multidimensional extensions of Theorem 6.1 on the basis of (5.8) are discussed in [48, 63], with connections with the classical Slepian inequality (cf. [53]).

Theorem 6.2 (Multidimensional Borell theorem). *Let $X = (X_1, \ldots, X_m)$ be a centered Gaussian vector in $\mathbb{R}^m$ with (non-degenerate) covariance matrix M such that $M_{k\ell} \geq 0$ for all $k, \ell = 1, \ldots, m$. Then, for any Borel sets $B_1, \ldots, B_m$ in $\mathbb{R}$,*

$$\mathbb{P}(X_1 \in B_1, \ldots, X_m \in B_m) \leq \mathbb{P}(X_1 \leq b_1, \ldots, X_m \leq b_m)$$

where $\mathbb{P}(X_k \in B_k) = \Phi(b_k/\sigma_k)$, $\sigma_k = \sqrt{M_{kk}}$, $k = 1, \ldots, m$.

References

[1] Ambrosio, L., Gigli, N., Savaré, G., *Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below*, Invent. Math. **195** (2014), 289–391.

[2] ———, *Metric measure spaces with Riemannian Ricci curvature bounded from below* (2012), Duke Math. J., to appear.

[3] ———, *Bakry-Émery curvature-dimension condition and Riemannian Ricci curvature bounds* (2012).

[4] Baernstein, A. II and Taylor, B. A., *Spherical rearrangements, subharmonic functions and *-functions in n-space*, Duke Math. J. **43** (1976), 245–268.

[5] Bakry, D., *Transformations de Riesz pour les semi-groupes symétriques. II. Étude sous la condition* $\Gamma_2 \geq 0$, Séminaire de Probabilités XIX, Lecture Notes in Math. 1123 (1985),145–174. Springer.

[6] ———, *L'hypercontractivité et son utilisation en théorie des semigroupes*, École d'Été de Probabilités de Saint-Flour, Lecture Notes in Math. 1581 (1994), 1–114. Springer.

[7] Bakry, D. and Émery, M., *Diffusions hypercontractives*, Séminaire de Probabilités XIX, Lecture Notes in Math. 1123 (1985), 177–206. Springer.

[8] Bakry, D., Gentil, I., and Ledoux, M., *On Harnack inequalities and optimal transportation* (2013), Ann. Scuola Norm. Pisa, to appear.

[9] ———, *Analysis and geometry of Markov diffusion operators*, Grundlehren der Mathematischen Wissenschaften 348, Springer, Berlin, 2014.

[10] Bakry, D. and Ledoux, M., *Lévy-Gromov's isoperimetric inequality for an infinite-dimensional diffusion generator*, Invent. Math. **123** (1996), 259–281.

[11] ———, *A logarithmic Sobolev form of the Li-Yau parabolic inequality*, Rev. Mat. Iberoam. **22** (2006), 683–702.

[12] Ball, K., *Volumes of sections of cubes and related problems*, Geometric aspects of functional analysis, Lecture Notes in Math. 1376 (1989), 251–260. Springer.

[13] Barthe, F., *On a reverse form of the Brascamp-Lieb inequality*, Invent. Math. **134** (1998), 335–361.

[14] Barthe, F. and Cordero-Erausquin, D., *Inverse Brascamp-Lieb inequalities along the heat equation*, Geometric aspects of functional analysis, Lecture Notes in Math. 1850 (2004), 65–71. Springer.

[15] Barthe, F., Cordero-Erausquin, D., Ledoux, M., and Maurey, B., *Correlation and Brascamp-Lieb inequalities for Markov semigroups*, Int. Math. Res. Not. IMRN **10** (2011), 2177–2216.

[16] Beckner, W., *Sobolev inequalities, the Poisson semigroup, and analysis on the sphere* $\mathbb{S}^n$, Proceedings of the National Academy of Sciences **89** (1992), 4816–4819.

[17] Bennett, J., *Heat-flow monotonicity related to some inequalities in euclidean analysis*, Proceedings of the 8th International Conference on Harmonic Analysis and Partial Differential Equations, Contemporary Mathematics 505 (2010), 85–96.

[18] Bennett, J., Bez, N., and Carbery, A., *Heat-flow monotonicity related to the Hausdorff-Young inequality*, Bull. Lond. Math. Soc. **41** (2009), 971–979.

[19] Bennett, J., Carbery, A., Christ, M., and Tao, T., *The Brascamp-Lieb inequalities: finiteness, structure and extremals*, Geom. Funct. Anal. **17** (2008), 1343–1415.

[20] ———, *On multilinear inequalities of Brascamp-Lieb type*, Math. Res. Lett. **17** (2010), 647–666.

[21] Bobkov, S., *An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in Gauss space*, Ann. Probab. **25** (1997), 206–214.

[22] Bobkov, S., Gentil, I., and Ledoux, M., *Hypercontractivity of Hamilton-Jacobi equations*, J. Math. Pures Appl. **80** (2001), 669–696.

[23] Borell, C., *The Brunn-Minkowski inequality in Gauss space*, Invent. Math. **30** (1975), 207–216.

[24] ———, *Geometric bounds on the Ornstein-Uhlenbeck velocity process*, Z. Wahrsch. Verw. Gebiete **70** (1985), 1–13.

[25] ———, *Diffusion equations and geometric inequalities*, Potential Anal. **12** (2000), 49–71.

[26] ———, *The Ehrhard inequality*, C. R. Math. Acad. Sci. Paris **337** (2003), 663–666.

[27] Brascamp, H. and Lieb, E., *Best constants in Young's inequality, its converse, and its generalization to more than three functions*, Advances in Math. **20** (1976), 151–173.

[28] Brascamp, H., Lieb, E., and Luttinger, J. M., *A general rearrangement inequality for multiple integrals*, J. Funct. Anal. **17** (1974), 227–237.

[29] Carbery, A., *The Brascamp-Lieb inequalities: recent developments*, Nonlinear Analysis, Function Spaces and Applications 8 (2007), 8–34. Czech. Acad. Sci., Prague.

[30] Carlen, E. and Loss, M., *Extremals of functionals with competing symmetries*, J. Funct. Analysis **88** (1990), 437–456.

[31] Carlen, E., Lieb, E., and Loss, M., *A sharp analog of Young's inequality on $\mathbb{S}^N$ and related entropy inequalities*, J. Geom. Anal. **14** (2004), 487–520.

[32] ———, *An inequality of Hadamard type for permanents*, Methods Appl. Anal. **13** (2006), 1–17.

[33] Carrillo, J., McCann, R., and Villani, C., *Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates*, Rev. Mat. Iberoam. **19** (2003), 971–1018.

[34] Chen W.-K., Dafnis, N., and Paouris, G., *Improved Hölder and reverse Hölder inequalities for correlated Gaussian random vectors* (2013).

[35] Cianchi, A., Fusco, N., Maggi F., and Pratelli, A., *On the isoperimetric deficit in Gauss space*, Amer. J. Math. **133** (2011), 131–186.

[36] Colding, T., *New monotonicity formulas for Ricci curvature and applications I*, Acta Math. **209** (2012), 229–263.

[37] Colding, T., Minicozzi, W. P. II, *Monotonicity – Analytic and geometric implications* (2012).

[38] Cordero-Erausquin, D. and Ledoux, M., *The geometry of Euclidean convolution inequalities and entropy*, Proc. Amer. Math. Soc. **138** (2010), 2755–2769.

[39] Davies, E. B., *Heat kernels and spectral theory*, Cambridge Tracts in Mathematics 92, Cambridge University Press, Cambridge, 1989.

[40] De, A., Mossel, E., and Neeman, J., *Majority is Stablest: Discrete and SoS* (2013).

[41] Ehrhard, A., *Symétrisation dans l'espace de Gauss*, Math. Scand. **53** (1983), 281–30.

[42] _____, *Inégalités isopérimétriques et intégrales de Dirichlet gaussiennes*, Ann. Sci. École Norm. Sup. **17** (1984), 317–332.

[43] Eldan, R., *A two-sided estimate for the Gaussian noise stability deficit* (2013).

[44] Erbar, M., Kuwada, K., and Sturm, K.-T., *On the equivalence of the entropy curvature-dimension condition and Bochner's inequality on metric measure spaces* (2013).

[45] Gallot, S., Hulin, D., and Lafontaine, J., *Riemannian geometry*, Third Edition, Universitext, Springer, Berlin, 2004.

[46] Gardner, R. J., *The Brunn-Minkowski inequality*, Bull. Amer. Math. Soc. **39** (2002), 355–405.

[47] Gross, L., *Logarithmic Sobolev inequalities*, Amer. J. Math. **97** (1975), 1061–1083.

[48] Isaksson, M. and Mossel, E., *New maximally stable Gaussian partitions with discrete applications*, Israel J. Math. **189** (2012), 347–396.

[49] Jordan, R., Kinderlehrer, D., and Otto, F., *The variational formulation of the Fokker-Planck equation*, SIAM J. Math. Anal. **29** (1998), 1–17.

[50] Kuwada, K., *Duality on gradient estimates and Wassertein controls*, J. Funct. Anal. **258** (2010), 3758–3774.

[51] Ledoux, M., *Isoperimetry and Gaussian analysis*, École d'Été de Probabilités de Saint-Flour, Lecture Notes in Math. 1648, (1996), 165–294. Springer.

[52] _____, *The concentration of measure phenomenon*, Mathematical Surveys and Monographs 89, American Mathematical Society, Providence, 2001.

[53] _____, *Remarks on Gaussian noise stability, Brascamp-Lieb and Slepian inequalities* (2014).

[54] Lieb, E. and Loss, M., *Analysis*, Second edition, Graduate Studies in Mathematics 14, American Mathematical Society, Providence, 2001.

[55] Li, P. and Yau, S.-T., *On the parabolic kernel of the Schrödinger operator*, Acta Math. **156** (1986), 153–201.

[56] Lichnerowicz, A., *Géométrie des groupes de transformations*, Travaux et Recherches Mathématiques III, Dunod, Paris, 1958.

[57] Lott, J. and Villani, C., *Ricci curvature for metric-measure spaces via optimal transport*, Ann. Math. **169** (2009), 903–991.

[58] Milman, E., *Sharp isoperimetric inequalities and model spaces for curvature-dimension-diameter condition* (2012), J. Eur. Math. Soc., to appear.

[59] Milman, V. and Schechtman, G., *Asymptotic theory of finite-dimensional normed spaces. With an appendix* by M. Gromov, Lecture Notes in Math. 1200, Springer, Berlin, 1986.

[60] Mossel, E. and Neeman, J., *Robust dimension free isoperimetry in Gaussian space* (2012), Ann. Probab., to appear.

[61] ———, *Robust optimality of Gaussian noise stability* (2012), J. Eur. Math. Soc., to appear.

[62] Mossel, E., O'Donnell, R., and Oleszkiewicz, K., *Noise stability of functions with low influences: invariance and optimality*, Ann. of Math. **171** (2010), 295–341.

[63] Neeman, J., *A multidimensional version of noise stability* (2013).

[64] Nelson, E., *The free Markoff field*, J. Funct. Anal. **12** (1973), 211–227.

[65] Ni, L., *The entropy formula for linear heat equation*, J. Geom. Anal. **14** (2004), 87–100.

[66] Otto, F., *The geometry of dissipative evolution equations: the porous medium equation*, Comm. Partial Differential Equations **26** (2001), 101–174.

[67] Otto, F. and Villani, C., *Generalization of an inequality by Talagrand, and links with the logarithmic Sobolev inequality*, J. Funct. Anal. **173** (2000), 361–400.

[68] Otto, F. and Westdickenberg, M., *Eulerian calculus for the contraction in the Wasserstein distance*, SIAM J. Math. Anal. **37** (2005), 1227–1255.

[69] von Renesse, M.-K. and Sturm, K.-T., *Transport inequalities, gradient estimates, entropy and Ricci curvature*, Comm. Pure Appl. Math. **68** (2005), 923–940.

[70] Riesz, F., *Sur une inégalité intégrale*, J. London Math. Soc. **5** (1930), 162–168.

[71] Rogers, C. A., *A single integral inequality*, J. London Math. Soc. **32** (1957), 102–108.

[72] Sturm, K.-T., *On the geometry of metric measure spaces* I & II, Acta Math. **196** (2006), 65–131 & 133–177.

[73] Sudakov, V. N. and Tsirel'son, B. S., *Extremal properties of half-spaces for spherically invariant measures*, J. Soviet. Math. **9** (1978), 9–18; translated from Zap. Nauch. Sem. L.O.M.I. **41** (1974), 14–24.

[74] Talagrand, M., *Transportation cost for Gaussian and other product measures*, Geom. Funct. Anal. **6** (1996), 587–600.

[75] Villani, C., *Topics in optimal transportation*, Graduate Studies in Mathematics 58, American Mathematical Society, Providence, 2003.

[76] ———, *Optimal transport. Old and new*, Grundlehren der Mathematischen Wissenschaften 338, Springer, Berlin, 2009.

[77] Wang, F.-Y., *Logarithmic Sobolev inequalities on noncompact Riemannian manifolds*, Probab. Theory Related Fields **109** (1997), 417–424.

[78] ———, *Functional inequalities, Markov properties and spectral theory*, Science Press, Beijing, 2005.

[79] ———, *Equivalent semigroup properties for the curvature-dimension condition*, Bull. Sci. Math. **135** (2011), 803–815.

Institut de Mathématiques de Toulouse, Université de Toulouse, F-31062 Toulouse, France, and Institut Universitaire de France

E-mail: ledoux@math.univ-toulouse.fr

Determinantal probability

Basic properties and conjectures

Russell Lyons

Abstract. We describe the fundamental constructions and properties of determinantal probability measures and point processes, giving streamlined proofs. We illustrate these with some important examples. We pose several general questions and conjectures.

Mathematics Subject Classification (2010). Primary 60K99, 60G55; Secondary 42C30, 37A15, 37A35, 37A50, 68U99.

Keywords. Random matrices, eigenvalues, orthogonal projections, positive contractions, exterior algebra, stochastic domination, negative association, point processes, mixtures, spanning trees, orthogonal polynomials, completeness, Bernoulli processes.

1. Introduction

Determinantal point processes were originally defined by Macchi [39] in physics. Starting in the 1990s, determinantal probability began to flourish as examples appeared in numerous parts of mathematics [51, 28, 8]. Recently, applications to machine learning have appeared [32].

A discrete determinantal probability measure is one whose elementary cylinder probabilities are given by determinants. More specifically, suppose that E is a finite or countable set and that Q is an $E \times E$ matrix. For a subset $A \subseteq E$, let $Q\restriction A$ denote the submatrix of Q whose rows and columns are indexed by A. If $\mathfrak{S}$ is a random subset of E with the property that for all finite $A \subseteq E$, we have

$$\mathbf{P}[A \subseteq \mathfrak{S}] = \det(Q\restriction A)\,, \tag{1.1}$$

then we call $\mathbf{P}$ a ***determinantal probability measure***. The inclusion-exclusion principle in combination with (1.1) yields the probability of each elementary cylinder event. Therefore, for every Q, there is at most one probability measure, to be denoted $\mathbf{P}^Q$, on subsets of E that satisfies (1.1). Conversely, it is known (see, e.g., [33]) that there is a determinantal probability measure corresponding to Q if Q is the matrix of a positive contraction on $\ell^2(E)$ (in the standard orthonormal basis).

Technicalities are required even to define the corresponding concept of determinantal point process for E being Euclidean space or a more general space. We present a virtually complete development of their basic properties in a way that minimizes such technicalities

Proceedings of International Congress of Mathematicians, 2014, Seoul

by adapting the approach of [33] from the discrete case. In addition, we use an idea of Goldman [21] to deduce properties of the general case from corresponding properties in the discrete case.

Space limitations prevent mention of most of what is known in determinantal probability theory, which pertains largely to the analysis of specific examples. We focus instead on some of the basic properties that hold for all determinantal processes and on some intriguing open questions.

2. Discrete basics

Let E be a denumerable set.

We identify a subset of E with an element of $\{0,1\}^E = 2^E$ in the usual way. There are several approaches to prove the basic existence results and identities for determinantal probability measures. We sketch the one used by [33]. This depends on understanding first the case where Q is the matrix of an orthogonal projection. It also relies on exterior algebra so that the existence becomes immediate.

Any unit vector v in a Hilbert space with orthonormal basis E gives a probability measure $\mathbf{P}^v$ on E, namely, $\mathbf{P}^v\big(\{e\}\big) := |(v,\, e)|^2$ for $e \in E$. Applying this simple idea to multivectors instead, we obtain the probability measures $\mathbf{P}^H$ associated to orthogonal projections P_H. We refer to [33] for details not given here.

2.1. Exterior algebra. Identify E with the standard orthonormal basis of the real or complex Hilbert space $\ell^2(E)$. For $k \geq 1$, let E_k denote a collection of ordered k-element subsets of E such that each k-element subset of E appears exactly once in E_k in some ordering. Define

$$\Lambda^k E := \bigwedge^k \ell^2(E) := \ell^2\Big(\big\{e_1 \wedge \cdots \wedge e_k\,;\ \langle e_1, \ldots, e_k\rangle \in E_k\big\}\Big)\,.$$

If $k > |E|$, then $E_k = \emptyset$ and $\Lambda^k E = \{0\}$. We also define $\Lambda^0 E$ to be the scalar field, $\mathbb{R}$ or $\mathbb{C}$. The elements of $\Lambda^k E$ are called ***multivectors*** of ***rank*** k, or ***k-vectors*** for short. We then define the ***exterior*** (or ***wedge***) ***product*** of multivectors in the usual alternating multilinear way: $\bigwedge_{i=1}^k e_{\sigma(i)} = (-1)^\sigma \bigwedge_{i=1}^k e_i$ for any permutation $\sigma \in \mathsf{Sym}(k)$, and

$$\bigwedge_{i=1}^k \sum_{e\in E'} a_i(e) e = \sum_{e_1,\ldots,e_k \in E'} \prod_{j=1}^k a_j(e_j) \bigwedge_{i=1}^k e_i$$

for any scalars $a_i(e)$ ($i \in [1,k],\ e \in E'$) and any finite $E' \subseteq E$. (Thus, $\bigwedge_{i=1}^k e_i = 0$ unless all e_i are distinct.) The inner product on $\Lambda^k E$ satisfies

$$(u_1 \wedge \cdots \wedge u_k,\, v_1 \wedge \cdots \wedge v_k) = \det\big[(u_i,\, v_j)\big]_{i,j\in[1,k]} \tag{2.1}$$

when u_i and v_j are 1-vectors. (This also shows that the inner product on $\Lambda^k E$ does not depend on the choice of orthonormal basis of $\ell^2(E)$.) We then define the ***exterior*** (or ***Grassmann***) ***algebra*** $\mathrm{Ext}\big(\ell^2(E)\big) := \mathrm{Ext}(E) := \bigoplus_{k\geq 0} \Lambda^k E$, where the summands are declared orthogonal, making it into a Hilbert space. Vectors $u_1, \ldots, u_k \in \ell^2(E)$ are linearly independent iff $u_1 \wedge \cdots \wedge u_k \neq 0$. For a k-element subset $A \subseteq E$ with ordering $\langle e_i\rangle$ in E_k, write $\theta_A := \bigwedge_{i=1}^k e_i$. We also write $\bigwedge_{e\in A} f(e) := \bigwedge_{i=1}^k f(e_i)$ for any function $f\colon E \to \ell^2(E)$.

Although there is an isometric isomorphism

$$u_1 \wedge \cdots \wedge u_k \mapsto \frac{1}{\sqrt{k!}} \sum_{\sigma \in \mathsf{Sym}(k)} (-1)^\sigma u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)} \in \ell^2(E^k)$$

for $u_i \in \ell^2(E)$, this does not simplify matters in the discrete case. It will be very useful in the continuous case later, however.

If H is a closed linear subspace of $\ell^2(E)$, written $H \le \ell^2(E)$, then we identify $\mathrm{Ext}(H)$ with its inclusion in $\mathrm{Ext}(E)$. That is, $\bigwedge^k H$ is the closure of the linear span of the k-vectors $\{v_1 \wedge \cdots \wedge v_k\,;\ v_1, \ldots, v_k \in H\}$. In particular, if $\dim H = r < \infty$, then $\bigwedge^r H$ is a 1-dimensional subspace of $\mathrm{Ext}(E)$; denote by ω_H a unit multivector in this subspace. Note that ω_H is unique up to a scalar factor of modulus 1; which scalar is chosen will not affect the definitions below. We denote by P_H the orthogonal projection onto H for any $H \le \ell^2(E)$ or, more generally, $H \le \mathrm{Ext}(E)$.

Lemma 2.1. *For every closed subspace $H \le \ell^2(E)$, every $k \ge 1$, and every $u_1, \ldots, u_k \in \ell^2(E)$, we have $P_{\mathrm{Ext}(H)}(u_1 \wedge \cdots \wedge u_k) = (P_H u_1) \wedge \cdots \wedge (P_H u_k)$.*

For $v \in \ell^2(E)$, write $[v]$ for the subspace of scalar multiples of v in $\ell^2(E)$.

2.2. Orthogonal projections. Let H be a subspace of $\ell^2(E)$ of dimension $r < \infty$. Define the probability measure $\mathbf{P}^H$ on subsets $B \subseteq E$ by

$$\mathbf{P}^H\big(\{B\}\big) := |\,(\omega_H,\ \theta_B)\,|^2\,. \tag{2.2}$$

Note that this is non-0 only for $|B| = r$. Also, by Lemma 2.1,

$$\mathbf{P}^H\big(\{B\}\big) = \|P_{\mathrm{Ext}(H)}\theta_B\|^2 = \|\bigwedge_{e\in B} P_H e\|^2$$

for $|B| = r$, which is non-0 iff $\langle P_H e\,;\ e \in B\rangle$ are linearly independent. In other words, $\mathbf{P}^H\big(\{B\}\big) \ne 0$ iff the projections of the elements of B form a basis of H. Let $\langle v_1, \ldots, v_r\rangle$ be any basis of H. If we use (2.1) and the fact that $\omega_H = c\bigwedge_i v_i$ for some scalar c, then we obtain another formula for $\mathbf{P}^H$:

$$\mathbf{P}^H\big(\{e_1, \ldots, e_r\}\big) = (\det[(v_i,\ e_j)]_{i,j\le r})^2 / \det[(v_i,\ v_j)]_{i,j\le r}\,. \tag{2.3}$$

We use $\mathfrak{B}$ to denote a random subset of E arising from a probability measure $\mathbf{P}^H$. To see that (1.1) holds for the matrix of P_H, observe that for $|B| = r$,

$$\mathbf{P}^H[\mathfrak{B} = B] = \big(P_{\mathrm{Ext}(H)}\theta_B,\ \theta_B\big) = \Big(\bigwedge_{e\in B} P_H e,\ \bigwedge_{e\in B} e\Big) = \det[(P_H e,\ f)]_{e,f\in B}$$

by (2.1). This shows that (1.1) holds for $|A| = r$ since $|\mathfrak{B}| = r$ $\mathbf{P}^H$-a.s. The general case is a consequence of multilinearity, which gives the following extension of (1.1). We use the convention that $\theta_\emptyset := 1$ and $\mathbf{u} \wedge 1 := \mathbf{u}$ for any multivector $\mathbf{u}$.

Theorem 2.2. *If A_1 and A_2 are (possibly empty) subsets of a finite set E, then*

$$\mathbf{P}^H[A_1 \subseteq \mathfrak{B}, A_2 \cap \mathfrak{B} = \emptyset] = \big(P_{\mathrm{Ext}(H)}\theta_{A_1} \wedge P_{\mathrm{Ext}(H^\perp)}\theta_{A_2},\ \theta_{A_1} \wedge \theta_{A_2}\big)\,. \tag{2.4}$$

In particular, for every $A \subseteq E$, we have

$$\mathbf{P}^H[A \subseteq \mathfrak{B}] = \|P_{\mathrm{Ext}(H)}\theta_A\|^2\,. \tag{2.5}$$

Corollary 2.3. *If E is finite, then for every subspace $H \le \ell^2(E)$, we have*

$$\forall B \subseteq E \quad \mathbf{P}^{H^\perp}\big(\{E \setminus B\}\big) = \mathbf{P}^{H}\big(\{B\}\big)\,. \tag{2.6}$$

These extend to infinite E. In order to define $\mathbf{P}^H$ when H is infinite dimensional, we proceed by finite approximation.

Let $E = \{e_i\,;\ i \ge 1\}$ be infinite. Consider first a finite-dimensional subspace H of $\ell^2(E)$. Define H_k as the image of the orthogonal projection of H onto the span of $\{e_i\,;\ 1 \le i \le k\}$. By considering a basis of H, we see that $P_{H_k} \to P_H$ in the weak operator topology (WOT), i.e., matrix-entrywise, as $k \to \infty$. It is also easy to see that if $r := \dim H$, then $\dim H_k = r$ for all large k and, in fact, $\omega_{H_k} \to \omega_H$ in the usual norm topology. It follows that (2.4) holds for this subspace H and for every finite $A_1, A_2 \subset E$.

Now let H be an infinite-dimensional closed subspace of $\ell^2(E)$. Choose finite-dimensional subspaces $H_k \uparrow H$. It is well known that $P_{H_k} \to P_H$ (WOT). Then

$$\text{for all finite sets } A \quad \det(P_{H_k}\restriction A) \to \det(P_H \restriction A)\,, \tag{2.7}$$

whence $\mathbf{P}^{H_k}$ has a weak* limit that we denote $\mathbf{P}^H$ and that satisfies (2.4).

We also note that for *any* sequence of subspaces H_k, if $P_{H_k} \to P_H$ (WOT), then $\mathbf{P}^{H_k} \to \mathbf{P}^H$ weak* because (2.7) then holds.

2.3. Positive contractions. We call Q a ***positive contraction*** if Q is a self-adjoint operator on $\ell^2(E)$ such that for all $u \in \ell^2(E)$, we have $0 \le (Qu, u) \le (u, u)$. A ***projection dilation*** of Q is an orthogonal projection P_H onto a closed subspace $H \le \ell^2(E')$ for some $E' \supseteq E$ such that for all $u \in \ell^2(E)$, we have $Qu = P_{\ell^2(E)} P_H u$, where we regard $\ell^2(E')$ as the orthogonal sum $\ell^2(E) \oplus \ell^2(E' \setminus E)$. In this case, Q is also called the ***compression*** of P_H to $\ell^2(E)$. Choose such a dilation (see (2.16) or (3.9)) and define $\mathbf{P}^Q$ as the law of $\mathfrak{B} \cap E$ when $\mathfrak{B}$ has the law $\mathbf{P}^H$. Then (1.1) for Q is a special case of (1.1) for P_H.

Of course, when Q is the orthogonal projection onto a subspace H, then $\mathbf{P}^Q = \mathbf{P}^H$. Basic properties of $\mathbf{P}^Q$ follow from those for orthogonal projections, such as:

Theorem 2.4. *If Q is a positive contraction, then for all finite $A_1, A_2 \subseteq E$,*

$$\mathbf{P}^Q\left[A_1 \subseteq \mathfrak{S}, A_2 \cap \mathfrak{S} = \emptyset\right] = \Big(\bigwedge_{e \in A_1} Qe \wedge \bigwedge_{e \in A_2} (I - Q)e,\ \theta_{A_1} \wedge \theta_{A_2} \Big)\,. \tag{2.8}$$

If (1.1) is given, then (2.8) can be deduced from (1.1) without using our general theory and, in fact, without assuming that the matrix Q is self-adjoint. Indeed, suppose that X is any diagonal matrix. Denote its (e,e)-entry by x_e. Comparing coefficients of x_e shows that (1.1) implies, for finite $A \subseteq E$,

$$\mathbf{E}\Big[\prod_{e \in A} \big(\mathbf{1}_{\{e \in \mathfrak{S}\}} + x_e\big) \Big] = \det\big((Q + X)\restriction A\big)\,. \tag{2.9}$$

Replacing A by $A_1 \cup A_2$ and choosing $x_e := -\mathbf{1}_{A_2}(e)$ gives (2.8). On the other hand, if we substitute $x_e := 1/(z_e - 1)$, then we may rewrite (2.9) as

$$\mathbf{E}\Big[\prod_{e \in A} \big(\mathbf{1}_{\{e \in \mathfrak{S}\}} z_e + \mathbf{1}_{\{e \notin \mathfrak{S}\}}\big) \Big] = \det\big((QZ + I - Q)\restriction A\big)\,, \tag{2.10}$$

where Z is the diagonal matrix of the variables z_e. Let E be finite. Write $z^A := \prod_{e\in A} z_e$ for $A \subseteq E$. Then (2.10) is equivalent to

$$\sum_{A\subseteq E} \mathbf{P}^Q[\mathfrak{S} = A] z^A = \det(I - Q + QZ)\,. \tag{2.11}$$

This is the same as the Laplace transform of $\mathbf{P}^Q$ after a trivial change of variables. When $\|Q\| < 1$, we can write $\det(I-Q+QZ) = \det(I-Q)\det(I+JZ)$ with $J := Q(I-Q)^{-1}$. Thus, for all $A \subseteq E$, we have

$$\mathbf{P}^Q[\mathfrak{S} = A] = \det(I - Q)\det(J\restriction A) = \det(I + J)^{-1}\det(J\restriction A)\,. \tag{2.12}$$

A probability measure $\mathbf{P}$ on 2^E is called ***strongly Rayleigh*** if its generating polynomial $f(z) := \sum_{A\subseteq E} \mathbf{P}[\mathfrak{S} = A] z^A$ satisfies the inequality

$$\frac{\partial f}{\partial z_e}(x)\frac{\partial f}{\partial z_{e'}}(x) \geq \frac{\partial^2 f}{\partial z_e \partial z_{e'}}(x) f(x) \tag{2.13}$$

for all $e \neq e' \in E$ and all real $x \in \mathbb{R}^E$. This property is satisfied by every determinantal probability measure, as was shown by [7], who demonstrated its usefulness in showing other properties, such as negative associations and preservation under symmetric exclusion processes.

For a set $K \subseteq E$, denote by $\mathscr{F}(K)$ the σ-field of events that are measurable with respect to the events $\{e \in \mathfrak{S}\}$ for $e \in K$. Define the ***tail*** σ-field to be the intersection of $\mathscr{F}(E \setminus K)$ over all finite K. We say that a measure $\mathbf{P}$ on 2^E has ***trivial tail*** if every event in the tail σ-field has measure either 0 or 1.

Theorem 2.5 ([33]). *If Q is a positive contraction, then $\mathbf{P}^Q$ has trivial tail.*

For finite E and a positive contraction Q, define the ***entropy*** of $\mathbf{P}^Q$ to be

$$\mathsf{Ent}(Q) := -\sum_{A\subseteq E} \mathbf{P}^Q(\{A\}) \log \mathbf{P}^Q(\{A\})\,.$$

Numerical calculation supports the following conjecture [33]:

Conjecture 2.6. *For all positive contractions Q_1 and Q_2, we have*

$$\mathsf{Ent}\big((Q_1 + Q_2)/2\big) \geq \big(\mathsf{Ent}(Q_1) + \mathsf{Ent}(Q_2)\big)/2\,. \tag{2.14}$$

2.4. Stochastic inequalities. Let E be denumerable. A function $f\colon 2^E \to \mathbb{R}$ is called ***increasing*** if for all $A \in 2^E$ and all $e \in E$, we have $f\big(A \cup \{e\}\big) \geq f(A)$. An event is called increasing or ***upwardly closed*** if its indicator is increasing.

Given two probability measures $\mathbf{P}^1, \mathbf{P}^2$ on 2^E, we say that $\mathbf{P}^2$ ***stochastically dominates*** $\mathbf{P}^1$ and write $\mathbf{P}^1 \preccurlyeq \mathbf{P}^2$ if for all increasing events $\mathcal{A}$, we have $\mathbf{P}^1(\mathcal{A}) \leq \mathbf{P}^2(\mathcal{A})$. This is equivalent to $\int f\, d\mathbf{P}^1 \leq \int f\, d\mathbf{P}^2$ for all bounded increasing f.

A ***coupling*** of two probability measures $\mathbf{P}^1$, $\mathbf{P}^2$ on 2^E is a probability measure μ on $2^E \times 2^E$ whose coordinate projections are $\mathbf{P}^1$, $\mathbf{P}^2$; it is ***monotone*** if

$$\mu\big\{(\mathcal{A}_1, \mathcal{A}_2)\,;\ \mathcal{A}_1 \subseteq \mathcal{A}_2\big\} = 1\,.$$

By Strassen's theorem [53], stochastic domination $\mathbf{P}^1 \preccurlyeq \mathbf{P}^2$ is equivalent to the existence of a monotone coupling of $\mathbf{P}^1$ and $\mathbf{P}^2$.

Theorem 2.7 ([33]). *If $H_1 \le H_2 \le \ell^2(E)$, then $\mathbf{P}^{H_1} \preccurlyeq \mathbf{P}^{H_2}$.*

It would be very interesting to find a natural or explicit monotone coupling.

A coupling μ has ***union marginal*** $\mathbf{P}$ if for all events $\mathcal{A} \subseteq 2^E$, we have $\mathbf{P}(\mathcal{A}) = \mu\{(A_1, A_2)\,;\ A_1 \cup A_2 \in \mathcal{A}\}$.

Question 2.8 ([33]). *Given $H = H_1 \oplus H_2$, is there a coupling of $\mathbf{P}^{H_1}$ and $\mathbf{P}^{H_2}$ with union marginal $\mathbf{P}^H$?*

A positive answer is supported by some numerical calculation. It is easily seen to hold when $H = \ell^2(E)$ by Corollary 2.3.

In the sequel, we write $Q_1 \preceq Q_2$ if $(Q_1 u,\, u) \le (Q_2 u,\, u)$ for all $u \in \ell^2(E)$.

Theorem 2.9 ([33, 7]). *If $0 \preceq Q_1 \preceq Q_2 \preceq I$, then $\mathbf{P}^{Q_1} \preccurlyeq \mathbf{P}^{Q_2}$.*

Proof. By Theorem 2.7, it suffices that there exist orthogonal projections P_1 and P_2 that are dilations of Q_1 and Q_2 such that $P_1 \preceq P_2$. This follows from Naĭmark's dilation theorem [43], which says that any measure whose values are positive operators, whose total mass is I, and which is countably additive in the weak operator topology dilates to a spectral measure. The measure in our case is defined on a 3-point space, with masses Q_1, $Q_2 - Q_1$, and $I - Q_2$, respectively. If we denote the respective dilations by R_1, R_2, and R_3, then we set $P_1 := R_1$ and $P_2 := R_1 + R_2$. □

A positive answer in general to Question 2.8 would give the following more general result by compression: If Q_1, Q_2 and $Q_1 + Q_2$ are positive contractions on $\ell^2(E)$, then there is a coupling of $\mathbf{P}^{Q_1}$ and $\mathbf{P}^{Q_2}$ with union marginal $\mathbf{P}^{Q_1+Q_2}$.

It would be very useful to have additional sufficient conditions for stochastic domination: see the end of Subsection 3.8 and Conjecture 5.7. For examples where more is known, see Theorem 5.2.

We shall say that the events in $\mathscr{F}(K)$ are ***measurable with respect to*** K and likewise for functions that are measurable with respect to $\mathscr{F}(K)$. We say that $\mathbf{P}$ has ***negative associations*** if for every pair f_1, f_2 of increasing functions that are measurable with respect to complementary subsets of E,

$$\mathbf{E}[f_1 f_2] \le \mathbf{E}[f_1]\mathbf{E}[f_2]\,. \tag{2.15}$$

Theorem 2.10 ([33]). *If $0 \preceq Q \preceq I$, then $\mathbf{P}^Q$ has negative associations.*

Proof. The details for finite E were given in [33]. For infinite E, let f_1 and f_2 be increasing bounded functions measurable with respect to $\mathscr{F}(A)$ and $\mathscr{F}(E \setminus A)$, respectively. Choose finite $E_n \uparrow E$. The conditional expectations $\mathbf{E}[f_1 \mid \mathscr{F}(A \cap E_n)]$ and $\mathbf{E}[f_2 \mid \mathscr{F}(E_n \setminus A)]$ are increasing functions to which (2.15) applies (because restriction to E_n corresponds to a compression of Q, which is a positive contraction) and which, being martingales, converge to f_1 and f_2 in $L^2(\mathbf{P}^Q)$. □

2.5. Mixtures. Write $\mathrm{Bern}(p)$ for the distribution of a Bernoulli random variable with expectation p. For $p_k \in [0,1]$, let $\mathrm{Bin}(\langle p_k \rangle)$ be the distribution of a sum of independent $\mathrm{Bern}(p_k)$ random variables. Recall that $[v]$ is the set of scalar multiples of v.

Theorem 2.11 ([1]; Lemma 3.4 of [48]; (2.38) of [49]; [26]). *Let Q be a positive contraction with spectral decomposition $Q = \sum_k \lambda_k P_{[v_k]}$, where $\langle v_k\,;\ k \ge 1\rangle$ are orthonormal. Let $I_k \sim \mathrm{Bern}(\lambda_k)$ be independent. Let $\mathfrak{H} := \bigoplus_k [I_k v_k]$; thus, $Q = \mathbf{E} P_{\mathfrak{H}}$. Then $\mathbf{P}^Q = \mathbf{E}\mathbf{P}^{\mathfrak{H}}$. Hence, if $\mathfrak{S} \sim \mathbf{P}^Q$, then $|\mathfrak{S}| \sim \mathrm{Bin}(\langle \lambda_k \rangle)$.*

Proof. By Theorem 2.9, it suffices to prove it when only finitely many $\lambda_k \neq 0$. Then by Theorem 2.4, we have $\mathbf{P}^Q[A \subseteq \mathfrak{S}] = \Big(\bigwedge_{e\in A} Qe,\ \theta_A\Big)$ for all $A \subseteq E$. Now

$$\bigwedge_{e\in A} Qe = \bigwedge_{e\in A} \sum_k \lambda_k P_{[v_k]} e = \sum_{j\colon A\to\mathbb{N}} \prod_{e\in A} \lambda_{j(e)} \bigwedge_{e\in A} P_{[v_{j(e)}]} e$$
$$= \sum_{j\colon A\rightarrowtail\mathbb{N}} \prod_{e\in A} \lambda_{j(e)} \bigwedge_{e\in A} P_{[v_{j(e)}]} e$$

because $v \wedge v = 0$ and $P_{[v]}e$ is a multiple of v, so none of the terms where j is not injective contribute. Thus,

$$\bigwedge_{e\in A} Qe = \sum_{j\colon A\rightarrowtail\mathbb{N}} \mathbf{E}\Big[\prod_{e\in A} I_{j(e)}\Big] \bigwedge_{e\in A} P_{[v_{j(e)}]} e = \mathbf{E}\Big[\sum_{j\colon A\rightarrowtail\mathbb{N}} \prod_{e\in A} I_{j(e)} \bigwedge_{e\in A} P_{[v_{j(e)}]} e\Big]$$
$$= \mathbf{E}\Big[\sum_{j\colon A\to\mathbb{N}} \prod_{e\in A} I_{j(e)} \bigwedge_{e\in A} P_{[v_{j(e)}]} e\Big] = \mathbf{E} \bigwedge_{e\in A} \sum_k I_k P_{[v_k]} e = \mathbf{E} \bigwedge_{e\in A} P_{\mathfrak{H}} e\,.$$

We conclude that $\mathbf{P}^Q[A \subseteq \mathfrak{S}] = \mathbf{E}\left(\bigwedge_{e\in A} P_{\mathfrak{H}} e,\ \theta_A\right) = \mathbf{E}\big[\mathbf{P}^{\mathfrak{H}}[A \subseteq \mathfrak{B}]\big]$ by (2.8). □

We sketch another proof: Let E' be disjoint from E with the same cardinality. Choose an orthonormal sequence $\langle v'_k\rangle$ in $\ell^2(E')$. Define

$$H := \bigoplus_k \big[\sqrt{\lambda_k} v_k + \sqrt{1-\lambda_k} v'_k\big] \leq \ell^2(E \cup E')\,. \tag{2.16}$$

Then Q is the compression of P_H to $\ell^2(E)$. Expanding $\omega_H = \bigwedge_k(\sqrt{\lambda_k} v_k + \sqrt{1-\lambda_k} v'_k)$ in the obvious way into orthogonal pieces and restricting to E, we obtain the desired equation from (2.2).

The first proof shows more generally the following: Let Q_0 be a positive contraction. Let $\langle v_k\,;\ k \geq 1\rangle$ be (not necessarily orthogonal) vectors such that $Q_0 + \sum_k P_{[v_k]} \preceq I$. Let I_k be independent Bernoulli random variables with $\mathbf{E}\sum_k I_k < \infty$. Write $\mathfrak{Q} := Q_0 + \sum_k I_k P_{[v_k]}$. Then $\mathbf{P}^{\mathbf{E}\mathfrak{Q}} = \mathbf{E}\mathbf{P}^{\mathfrak{Q}}$. This was observed by Ghosh and Krishnapur (personal communication, 2014).

Note that in the mixture of Theorem 2.11, the distribution of $\langle I_k\,;\ k \geq 1\rangle$ is determinantal corresponding to the diagonal matrix with diagonal $\langle \lambda_k\,;\ k \geq 1\rangle$. Thus, it is natural to wonder whether $\langle I_k\,;\ k \geq 1\rangle$ can be taken to be a general determinantal measure. If such a mixture is not necessarily determinantal, must it be strongly Rayleigh or at least have negative correlations? Here, we say that a probability measure $\mathbf{P}$ on 2^E has ***negative correlations*** if for every pair A, B of finite disjoint subsets of E, we have $\mathbf{P}[A \cup B \subseteq \mathfrak{S}] \leq \mathbf{P}[A \subseteq \mathfrak{S}]\mathbf{P}[B \subseteq \mathfrak{S}]$. Note that negative associations is stronger than negative correlations.

2.6. Example: Uniform spanning trees and forests. The most well-known example of a (nontrivial discrete) determinantal probability measure is that where $\mathfrak{S}$ is a uniformly chosen random spanning tree of a finite connected graph $G = (\mathsf{V}, \mathsf{E})$ with $E := \mathsf{E}$. Here, we regard a spanning tree as a set of edges. The fact that (1.1) holds for the uniform spanning tree is due to [12] and is called the Transfer Current Theorem. The case with $|A| = 1$ was shown much earlier by [30], while the case with $|A| = 2$ was first shown by [11]. Write UST_G for the uniform spanning tree measure on G.

To see that UST_G is indeed determinantal, consider the vertex-edge incidence matrix M of G, where each edge is oriented (arbitrarily) and the (x,e)-entry of M equals 1 if x is the head of e, -1 if x is the tail of e, and 0 otherwise. Identifying an edge with its corresponding column of M, we find that a spanning tree is the same as a basis of the column space of M. Given $x \in \mathsf{V}$, define the ***star*** at x to be the x-row of M, regarded as a vector $\star_x$ in the row space, $\bigstar(G) \le \ell^2(\mathsf{E})$. It is easy that the row-rank of M is $|\mathsf{V}| - 1$. Let $x_0 \in \mathsf{V}$ and let $\mathbf{u}$ be the wedge product (in some order) of the stars at all the vertices other than x_0. Thus, $\mathbf{u} = c\,\omega_{\bigstar(G)}$ for some $c \neq 0$. Since spanning trees are bases of the column space of M, we have $\big(\mathbf{u},\ \theta_A\big) \neq 0$ iff A is a spanning tree. That is, the only non-zero coefficients of $\mathbf{u}$ are those in which choosing one edge in each $\star_x$ for $x \neq x_0$ yields a spanning tree; moreover, each spanning tree occurs exactly once since there is exactly one way to choose an edge incident to each $x \neq x_0$ to get a given spanning tree. This means that its coefficient is ± 1. Hence, $\mathbf{P}^{\bigstar(G)}$ is indeed uniform on spanning trees. Simultaneously, this proves the matrix tree theorem that the number of spanning trees equals $\det[(\star_x,\, \star_y)]_{x,y\neq x_0}$, since this determinant is $\|\mathbf{u}\|^2$.

One can define analogues of UST_G on infinite connected graphs [44, 22, 2] by weak limits. For brevity, we simply define them here as determinantal probability measures. Again, all edges of G are oriented arbitrarily. We define $\bigstar(G)$ as the closure of the linear span of the stars. An element of $\ell^2(\mathsf{E})$ that is finitely supported and orthogonal to $\bigstar(G)$ is called a ***cycle***; the closed linear span of the cycles is $\diamondsuit(G)$. The ***wired uniform spanning forest*** is $\mathsf{WSF}_G := \mathbf{P}^{\bigstar(G)}$, while the ***free uniform spanning forest*** is $\mathsf{FSF}_G := \mathbf{P}^{\diamondsuit(G)^\perp}$.

3. Continuous basics

Our discussion of the "continuous" case includes the discrete case, but the discrete case has the more elementary formulations given earlier.

Let E be a measurable space. As before, E will play the role of the underlying set on which a point process forms a counting measure. While before we implicitly used counting measure on E itself, now we shall have an arbitrary measure μ; it need not be a probability measure. The case of Lebesgue measure on Euclidean space is a common one. The Hilbert spaces of interest will be $L^2(E,\mu)$.

3.1. Symmetrization and anti-symmetrization. There may be no natural order in E, so to define, e.g., a probability measure on n points of E, it is natural to use a probability measure on E^n that is symmetric under coordinate changes and that vanishes on the diagonal $\Delta_n(E) := \big\{(x_1,\dots,x_n) \in E^n\,;\ \exists i \neq j\ \ x_i = x_j\big\}$. Likewise, for exterior algebra, it is more convenient to identify $u_1 \wedge \cdots \wedge u_n$ with

$$\sum_{\sigma\in\mathsf{Sym}(n)} (-1)^\sigma u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}/\sqrt{n!} \in L^2(E^n,\mu^n)$$

for $u_i \in L^2(E,\mu)$. Thus, $u_1 \wedge \cdots \wedge u_n$ is identified with the function

$$(x_1,\dots,x_n) \mapsto \det[u_i(x_j)]_{i,j\in\{1,\dots,n\}}/\sqrt{n!}\,.$$

Note that

$$n!\Big(\bigwedge_{i=1}^n u_i\Big)\Big(\bigwedge_{i=1}^n v_i\Big)(x_1,\dots,x_n) = \det[u_i(x_j)]\det[v_i(x_j)] = \det[u_i(x_j)]\det[v_i(x_j)]^T$$
$$= \det[u_i(x_j)][v_i(x_j)]^T = \det[K(x_i,x_j)]_{i,j\in\{1,\dots,n\}} \tag{3.1}$$

with $K := \sum_{i=1}^n u_i \otimes v_i$. Here, ^{T} denotes transpose.

3.2. Joint intensities. Suppose from now on that E is a locally compact Polish space (equivalently, a locally compact second countable Hausdorff space). Let μ be a Radon measure on E, i.e., a Borel measure that is finite on compact sets. Let $\mathcal{N}(E)$ be the set of Radon measures on E with values in $\mathbb{N}\cup\{\infty\}$. We give $\mathcal{N}(E)$ the vague topology generated by the maps $\xi \mapsto \int f\,d\xi$ for continuous f with compact support; then $\mathcal{N}(E)$ is Polish. The corresponding Borel σ-field of $\mathcal{N}(E)$ is generated by the maps $\xi\mapsto\xi(A)$ for Borel $A\subseteq E$. Let $\mathfrak{X}$ be a simple point process on E, i.e., a random variable with values in $\mathcal{N}(E)$ such that $\mathfrak{X}(\{x\})\in\{0,1\}$ for all $x\in E$. The power $\mathfrak{X}^k := \mathfrak{X}\otimes\cdots\otimes\mathfrak{X}$ lies in $\mathcal{N}(E^k)$. Thus, $\mathbf{E}[\mathfrak{X}^k]$ is a Borel measure on E^k; the part of it that is concentrated on $E^k\setminus\Delta_k(E)$ is called the ***k*-point intensity measure** of $\mathfrak{X}$. If the intensity measure is absolutely continuous with respect to μ^k, then its Radon-Nikodym derivative ρ_k is called the ***k*-point intensity function** or the ***k*-point correlation function**:

$$\text{for all Borel } A\subseteq E^k\setminus\Delta_k(E) \qquad \mathbf{E}[\mathfrak{X}^k(A)] = \int_A \rho_k\,d\mu^k\,. \tag{3.2}$$

Since the intensity measure vanishes on the diagonal $\Delta_k(E)$, we take ρ_k to vanish on $\Delta_k(E)$. We also take ρ_k to be symmetric under permutations of coordinates. Intensity functions are the continuous analogue of the elementary probabilities (1.1).

Since the sets $\prod_{i=1}^k A_i := A_1\times\cdots\times A_k$ generate the σ-field on $E^k\setminus\Delta_k(E)$ for pairwise disjoint Borel $A_1,\dots,A_k\subseteq E$, a measurable function $\rho_k\colon E^k\to[0,\infty)$ is "the" k-point intensity function iff

$$\mathbf{E}\Big[\prod_{i=1}^k \mathfrak{X}(A_i)\Big] = \int_{\prod_{i=1}^k A_i} \rho_k\,d\mu^k\,. \tag{3.3}$$

Since $\mathfrak{X}$ is simple, $\mathfrak{X}^k\big(A^k\setminus D_k(A)\big) = \big(\mathfrak{X}(A)\big)_k$, where $(n)_k := n(n-1)\cdots(n-k+1)$. Since ρ_k vanishes on the diagonal, it follows from (3.2) that for disjoint $A_1,\dots,A_r$ and non-negative $k_1,\dots,k_r$ summing to k,

$$\mathbf{E}\Big[\prod_{j=1}^r \big(\mathfrak{X}(A_j)\big)_{k_j}\Big] = \int_{\prod_{j=1}^r A_j^{k_j}} \rho_k\,d\mu^k\,. \tag{3.4}$$

Again, this characterizes ρ_k, even if we use only $r=1$.

In the special case that $\mathfrak{X}(E)=n$ a.s. for some $n\in\mathbb{Z}^+$, then the definition (3.2) shows that a random ordering of the n points of $\mathfrak{X}$ has density $\rho_n/n!$. More generally, (3.2) shows that for all $k<n$,

$$\text{the density of a random (ordered) } k\text{-tuple of } \mathfrak{X} \text{ is } \rho_k/(n)_k\,, \tag{3.5}$$

whence in this case,

$$\rho_k(x_1,\dots,x_k) = \frac{1}{(n-k)!}\int_{E^{n-k}} \rho_n(x_1,\dots,x_n)\,d\mu^{n-k}(x_{k+1},\dots,x_n)\,. \tag{3.6}$$

We call $\mathfrak{X}$ ***determinantal*** if for some measurable $K\colon E^2 \to \mathbb{C}$ and all $k \ge 1$, $\rho_k(F) = \det(K\restriction F)$ μ^k-a.e. Here, $K\restriction(x_1,\dots,x_k)$ is the matrix $[K(x_i,x_j)]_{i,j\le k}$. In this case, we denote the law of $\mathfrak{X}$ by $\mathbf{P}^K$.

We consider only K that are locally square integrable (i.e., $|K|^2\mu^2$ is Radon), are Hermitian (i.e., $K(y,x) = \overline{K(x,y)}$ for all $x,y \in E$), and are positive semidefinite (i.e., $K\restriction F$ is positive semidefinite for all finite F, written $K \succeq 0$). In this case, K defines a positive semidefinite integral operator $(Kf)(x) := \int K(x,y)f(y)\,d\mu(y)$ on functions $f \in L^2(\mu)$ with compact support. For every Borel $A \subseteq E$, we denote by μ_A the measure μ restricted to Borel subsets of A and by K_A the compression of K to A, i.e., $K_A f := (Kf)\restriction A$ for $f \in L^2(A,\mu_A)$. The operator K is locally trace-class, i.e., for every compact $A \subseteq E$, the compression K_A is trace class, having a spectral decomposition $K_A f = \sum_k \lambda_k^A (f,\,\phi_k^A)\phi_k^A$, where $\langle \phi_k^A\,;\ k \ge 1\rangle$ are orthonormal eigenfunctions of K_A with positive summable eigenvalues $\langle \lambda_k^A\,;\ k \ge 1\rangle$. If A_1 is the set where $\sum_k \lambda_k^A |\phi_k^A|^2 < \infty$, then $\mu(A \setminus A_1) = 0$ and $\sum_k \lambda_k^A \phi_k^A \otimes \overline{\phi_k^A}$ converges on A_1^2, with sum μ_A^2-a.e. equal to K. We normally redefine K on a set of measure 0 to equal this sum. Such a K defines a determinantal point process iff the integral operator K extends to all of $L^2(\mu)$ as a positive contraction [39, 51, 26]. The joint intensities determine uniquely the law of the point process [27, Lemma 4.2.6]. Poisson processes are not determinantal processes, but when μ is continuous, they are distributional limits of determinantal processes.

3.3. Construction. To see that a positive contraction defines a determinantal point process, we first consider K that defines an orthogonal projection onto a finite-dimensional subspace, H. Then $K = \sum_{k=1}^n \phi_k \otimes \overline{\phi_k}$ for every orthonormal basis $\langle \phi_k\,;\ k \le n\rangle$ of H and $\omega_H = \bigwedge_{i=1}^n \phi_k$ is a unit multivector in the notation of Subsection 2.1. Because of (3.1), we have

$$\frac{1}{n!}\int \det[K(x_i,x_j)]_{i,j\le n}\,d\mu^n(x_1,\dots,x_n) = \Big\|\bigwedge_{k=1}^n \phi_k\Big\|^2 = 1\,, \tag{3.7}$$

i.e., $\det[K(x_i,x_j)]/n!$ is a density with respect to μ^n. Although in the discrete case, the absolute squared coefficients of $\bigwedge_{k=1}^n \phi_k$ give the elementary probabilities, now coefficients are replaced by a function whose absolute square gives a probability density. As noted already, (3.7) means that $F \mapsto \det(K\restriction F)$ is the n-point intensity function. In order to show that this density gives a determinantal process with kernel K, we use the Cauchy-Binet formula, which may be stated as follows: For $k \times n$ matrices $a = [a_{i,j}]$ and $b = [b_{i,j}]$ with $a^J := [a_{i,j}]_{\substack{i\le k\\ j\in J}}$, we have

$$\det\left([a_{i,j}][b_{i,j}]^T\right) = \sum_{|J|=k} \det a^J \cdot \det b^J = \sum_{\substack{\sigma,\tau\in \mathsf{Sym}(k,n)\\ \mathrm{im}(\sigma)=\mathrm{im}(\tau)}} (-1)^\sigma (-1)^\tau \prod_{i=1}^k a_{i,\sigma(i)} b_{i,\tau(i)}\,,$$

where $\mathrm{im}(\sigma)$ denotes the image of σ and the sums extend over all pairs of injections

$$\sigma,\tau\colon \{1,2,\dots,k\} \rightarrowtail \{1,2,\dots,n\}\,.$$

Here, the sign $(-1)^\sigma$ of σ is defined in the usual way by the parity of the number of pairs $i<j$ for which $\sigma(i)>\sigma(j)$. We have

$$\begin{aligned}\rho_k(x_1,\dots,x_k) &= \frac{1}{(n-k)!}\int_{E^{n-k}} \det[K(x_i,x_j)]\,d\mu^{n-k}(x_{k+1},\dots,x_n)\\ &= \frac{1}{(n-k)!}\int_{E^{n-k}} \sum_{\sigma\in\mathsf{Sym}(n)}(-1)^\sigma\prod_{i=1}^n \phi_{\sigma(i)}(x_i)\cdot\\ &\qquad\cdot\sum_{\tau\in\mathsf{Sym}(n)}(-1)^\tau\prod_{i=1}^n\overline{\phi_{\tau(i)}(x_i)}\,d\mu^{n-k}(x_{k+1},\dots,x_n) \qquad (3.8)\\ &= \sum_{\substack{\sigma,\tau\in\mathsf{Sym}(k,n)\\ \mathrm{im}(\sigma)=\mathrm{im}(\tau)}}(-1)^\sigma(-1)^\tau\prod_{i=1}^k\phi_{\sigma(i)}(x_i)\overline{\phi_{\tau(i)}(x_i)}\\ &= \det\big(K\restriction(x_1,\dots,x_k)\big)\,.\end{aligned}$$

Here, the first equality uses (3.6), the second equality uses (3.1), the third equality uses the fact that $\int_E \phi_{\sigma(i)}(x_i)\overline{\phi_{\tau(i)}(x_i)}\,d\mu(x_i)$ is 1 or 0 according as $\sigma(i)=\tau(i)$ or not, and the fourth equality uses Cauchy-Binet. Note that a factor of $(n-k)!$ arises because for every pair of injections $\sigma_1,\tau_1\in\mathsf{Sym}(k,n)$ with equal image, there are $(n-k)!$ extensions of them to permutations $\sigma,\tau\in\mathsf{Sym}(n)$ with $\sigma(i)=\tau(i)$ for all $i>k$; in this case, $(-1)^\sigma(-1)^\tau=(-1)^{\sigma_1}(-1)^{\tau_1}$. We write $\mathbf{P}^H$ for the law of the associated point process on E.

Lemma 3.1. *Let $\mathfrak{X}_n\sim\mathbf{P}^{K_n}$ with $K_n(x,x)\le f(x)$ for some $f\in L^1_{\mathrm{loc}}(E,\mu)$. Then $\{\mathbf{P}^{K_n}\,;\,n\ge1\}$ is tight and every weak limit point of $\mathfrak{X}_n$ is simple.*

Proof. By using the kernel $K_n(x,y)/\sqrt{f(x)f(y)}$ with respect to the measure $f\mu$, we may assume that $f\equiv1$. Tightness follows from

$$m\mathbf{P}[\mathfrak{X}_n(A)\ge m]\le\mathbf{E}[\mathfrak{X}_n(A)]=\int_A K_n(x,x)\,d\mu(x)\,.$$

For the rest, we may assume that E is compact and $\mu(E)=1$. Let $\mathfrak{X}$ be a limit point of $\mathfrak{X}_n$. Let μ_{d} be the atomic part of μ and $\mu_{\mathrm{c}}:=\mu-\mu_{\mathrm{d}}$. Choose $m\ge1$ and partition E into sets $A_1,\dots,A_m$ with $\mu_{\mathrm{c}}(A_i)\le1/m$. Let A be such that $\mu_{\mathrm{d}}(E\setminus A)=0$ and $\mu_{\mathrm{c}}(A)=0$. Let U be open such that $A\subseteq U$ and $\mu_{\mathrm{c}}(U)<1/m$. Then

$$\begin{aligned}\mathbf{P}[\mathfrak{X}\text{ is not simple}] &\le \limsup_n\big(\mathbf{P}[\mathfrak{X}_n(U\setminus A)\ge1]+\mathbf{P}[\exists i\ \ \mathfrak{X}_n(A_i)\ge2]\big)\\ &\le \limsup_n\big(\mathbf{E}[\mathfrak{X}_n(U\setminus A)]+\sum_i\mathbf{E}[(\mathfrak{X}_n(A_i))_2]\big)\\ &\le \mu_{\mathrm{c}}(U)+\sum_i\mu(A_i)^2<2/m\,.\end{aligned}$$

□

Now, given any locally trace-class orthogonal projection K onto H, choose finite-dimensional subspaces $H_n\uparrow H$ with corresponding projections K_n. Clearly $K_n(x,y)\to K(x,y)$ μ^2-a.e. and $K_n(x,x)\le K(x,x)$ μ-a.e. Thus, the joint intensity functions converge a.e.

By dominated convergence, if $A \subset E^k \setminus \Delta_k(E)$ is relatively compact and Borel, then $\mathbf{E}^{H_n}[\mathfrak{X}(A)] \to \int_A \det(K\restriction F)\, d\mu^k(F)$. By uniform exponential moments of $\mathfrak{X}(A)$ [27, proof of Lemma 4.2.6], it follows that all weak limit points of $\mathbf{P}^{H_n}$ are equal, and hence, by Lemma 3.1, define $\mathbf{P}^H$ with kernel K. (In Subsection 3.7, we shall see that $\langle \mathbf{P}^{H_n}\,;\ n \ge 1\rangle$ is stochastically increasing.)

Finally, let K be any locally trace-class positive contraction. Define the orthogonal projection on $L^2(E,\mu) \oplus L^2(E,\mu)$ whose block matrix is

$$\begin{pmatrix} K & \sqrt{K(I-K)} \\ \sqrt{K(I-K)} & I-K \end{pmatrix}. \tag{3.9}$$

Take an isometric isomorphism of $L^2(E,\mu)$ to $\ell^2(E')$ for some denumerable set E' and interpret the above as an orthogonal projection K' on $L^2(E,\mu)\oplus\ell^2(E')$. Then K' is clearly locally trace-class and K is the compression of K' to E. Thus, we define $\mathbf{P}^K$ by intersecting samples of $\mathbf{P}^{K'}$ with E. We remark that by writing K' as a limit of increasing finite-rank projections that we then compress, we see that $\mathbf{P}^K$ may be defined as a limit of determinantal processes corresponding to increasing finite-rank positive contractions.

Conjecture 3.2. *If K is a locally trace-class positive contraction, then $\mathbf{P}^K$ has trivial tail in that every event in $\bigcap_{\text{compact } A\subset E} \mathscr{F}(E\setminus A)$ is trivial.*

3.4. Mixtures. Rather than using compressions as in the last paragraph above, an alternative approach to defining $\mathbf{P}^K$ uses mixtures and starts from finite-rank projections, as in Subsection 2.5. This approach is due to [26]. Consider first a finite-rank $K := \sum_{j=1}^n \lambda_j \phi_j \otimes \overline{\phi_j}$. Let $I_j \sim \mathrm{Bern}(\lambda_j)$ be independent. Let $\mathfrak{H} := \bigoplus_j [I_j\phi_j]$; thus, $K = \mathbf{E}P_{\mathfrak{H}}$. We claim that $\mathbf{P}^K := \mathbf{E}\mathbf{P}^{\mathfrak{H}}$ is determinantal with kernel K. Indeed, it is clearly a simple point process. Write $\Phi_J := \bigwedge_{j\in J} I_j\phi_j$, $\psi_j := \sqrt{\lambda_j}\phi_j$, and $\psi_J := \bigwedge_{j\in J}\psi_j$. Let $F \in E^k$. Combining Cauchy-Binet with (3.1) yields $\det\big(K\restriction F\big) = k!\sum_{|J|=k} |\psi_J(F)|^2$. Similarly, the joint intensities of $\mathbf{E}\mathbf{P}^{\mathfrak{H}}$ are the expectations of the joint intensities of $\mathbf{P}^{\mathfrak{H}}$, which equal

$$\mathbf{E}\big[\det(P_{\mathfrak{H}}\restriction F)\big] = \mathbf{E}\Big[k!\sum_{|J|=k} |\Phi_J(F)|^2\Big] = \det\big(K\restriction F\big).$$

Essentially the same works for trace-class $K = \sum_{j=1}^\infty \lambda_j\phi_j\otimes\overline{\phi_j}$; we need merely take, in the last step, a limit in the above equation as $n\to\infty$ for $K_n := \sum_{j=1}^n \lambda_j\phi_j\otimes\overline{\phi_j}$, since all terms are non-negative and $K_n \to K$ a.e.

Given this construction of $\mathbf{P}^K$ for trace-class K, one can then construct $\mathbf{P}^K$ for a general locally trace-class positive contraction by defining its restriction to each relatively compact set A via the trace-class compression K_A.

As noted by [26], a consequence of the mixture representation is a CLT due originally to [52]:

Theorem 3.3. *Let K_n be trace-class positive contractions on spaces $L^2(E_n,\mu_n)$. Let $\mathfrak{X}_n \sim \mathbf{P}^{K_n}$ and write $|\mathfrak{X}_n| := \mathfrak{X}_n(E_n)$. If $\mathrm{Var}\big(|\mathfrak{X}_n|\big) \to \infty$ as $n\to\infty$, then $\langle |\mathfrak{X}_n|\,;\ n\ge 1\rangle$ obeys a CLT.*

3.5. Simulation. In order to simulate $\mathbf{P}^K$ when K is a trace-class positive contraction, it suffices, by taking a mixture as above, to see how to simulate $\mathfrak{X} \sim \mathbf{P}^H$ when $n :=$

$\dim H < \infty$. The following algorithm [26, Algo. 18] gives a uniform random ordering of $\mathfrak{X}$ as $\langle X_1, \dots, X_n\rangle$. Since $\mathbf{E}\big[\mathfrak{X}(E)\big] = n$, the measure $\mathbf{E}[\mathfrak{X}]/n = n^{-1}K(x,x)\,d\mu(x)$ is a probability measure on E. Select a point X_1 at random from that measure. If $n = 1$, then we are done. If not, then let H_1 be the orthogonal complement in H of the function $K_{X_1} := \sum_{k=1}^n \overline{\phi_k(X_1)}\phi_k \in H$, where $\langle \phi_k\,;\ k \le n\rangle$ is an orthonormal basis for H. Then $\dim H_1 = n-1$ and we may repeat the above for H_1 to get the next point, X_2, then $H_2 := H_1 \cap K_{X_2}^{\perp}$, etc. The conditional density of X_{k+1} given $X_1, \dots, X_k$ is $(n-k)^{-1}\det\big(K\restriction(x, X_1, \dots, X_k)\big)/\det\big(K\restriction(X_1, \dots, X_k)\big)$ by (3.5), i.e., $(n-k)^{-1}$ times the squared distance from K_x to the linear span of $K_{X_1}, \dots, K_{X_k}$. It can help for rejection sampling to note that this is at most $(n-k)^{-1}K(x,x)$. One can also sample faster by noting that the conditional distribution of X_{k+1} is the same as that of $\mathbf{P}^{\mathfrak{v}}$, where $\mathfrak{v}$ is a uniformly random vector on the unit sphere of H_k.

3.6. Transference principle. Note that if $N_1, \dots, N_r$ are bounded $\mathbb{N}$-valued random variables, then the function $(k_1, \dots, k_r) \mapsto \mathbf{E}\Big[\prod_{j=1}^r \big(N_j\big)_{k_j}\Big]$ determines the joint distribution of $\langle N_j\,;\ j \le r\rangle$ since it gives the derivatives at $(1, 1, \dots, 1)$ of the probability generating function $(s_1, \dots, s_r) \mapsto \mathbf{E}\Big[\prod_{j=1}^r s_j^{N_j}\Big]$.

Let us re-examine (3.4) in the context of a finite-rank $K = \sum_{i=1}^n \lambda_i \phi_i \otimes \overline{\phi_i}$. Given disjoint $A_1, \dots, A_r \subseteq E$ and non-negative $k_1, \dots, k_r$ summing to k, it will be convenient to write $\kappa(j) := \min\big\{m \ge 1\,;\ j \le \sum_{\ell=1}^m k_\ell\big\}$ for $j \le k$. We have by Cauchy-Binet

$$\begin{aligned}
\mathbf{E}^K\Big[\prod_{\ell=1}^r \big(\mathfrak{X}(A_\ell)\big)_{k_\ell}\Big] &= \int_{\prod_{\ell=1}^r A_\ell^{k_\ell}} \rho_k\, d\mu^k = \int_{\prod_{\ell=1}^r A_\ell^{k_\ell}} \det(K\restriction(x_1, \dots, x_k)) \prod_{j=1}^k d\mu(x_j) \\
&= \int_{\prod_{\ell=1}^r A_\ell^{k_\ell}} \sum_{\substack{\sigma,\tau\in\mathsf{Sym}(k,n)\\ \mathrm{im}(\sigma)=\mathrm{im}(\tau)}} (-1)^\sigma(-1)^\tau \prod_{j=1}^k \lambda_{\sigma(j)}\phi_{\sigma(j)}(x_j)\overline{\phi_{\tau(j)}(x_j)} \prod_{j=1}^k d\mu(x_j) \\
&= \sum_{\substack{\sigma,\tau\in\mathsf{Sym}(k,n)\\ \mathrm{im}(\sigma)=\mathrm{im}(\tau)}} (-1)^\sigma(-1)^\tau \prod_{j=1}^k \int_{A_{\kappa(j)}} \lambda_{\sigma(j)}\phi_{\sigma(j)}(x_j)\overline{\phi_{\tau(j)}(x_j)}\, d\mu(x_j) \\
&= \sum_{\substack{\sigma,\tau\in\mathsf{Sym}(k,n)\\ \mathrm{im}(\sigma)=\mathrm{im}(\tau)}} (-1)^\sigma(-1)^\tau \lambda^{\mathrm{im}(\sigma)} \prod_{j=1}^k \big(\mathbf{1}_{A_{\kappa(j)}}\phi_{\sigma(j)},\ \overline{\phi_{\tau(j)}}\big) \\
&= \sum_{\sigma\in\mathsf{Sym}(k,n)} (-1)^\sigma \lambda^{\mathrm{im}(\sigma)} \det\Big[\Big(\mathbf{1}_{A_{\kappa(j)}}\phi_{\sigma(j)},\ \overline{\phi_\ell}\Big)\Big]_{\substack{j\le k\\ \ell\in\mathrm{im}(\sigma)}}.
\end{aligned}$$

As an immediate consequence of this formula, we obtain the following important principle of Goldman [21, Proposition 12] that allows one to infer properties of continuous determinantal point processes from corresponding properties of discrete determinantal probability measures:

Theorem 3.4. *Let (E,μ) and (F,ν) be two Radon measure spaces on locally compact Polish sets. Let $\langle A_i\rangle$ be pairwise disjoint Borel subsets of E and $\langle B_i\rangle$ be pairwise disjoint Borel subsets of F. Let $\lambda_k \in [0,1]$ with $\sum_k \lambda_k < \infty$. Let $\langle \phi_k\rangle$ be orthonormal in $L^2(E,\mu)$ and $\langle \psi_k\rangle$ be orthonormal in $L^2(F,\nu)$. Let $K := \sum_k \lambda_k \phi_k \otimes \overline{\phi_k}$ and $L := \sum_k \lambda_k \psi_k \otimes \overline{\psi_k}$.*

If $\big(\mathbf{1}_{A_i}\phi_j,\ \phi_k\big) = \big(\mathbf{1}_{B_i}\psi_j,\ \psi_k\big)$ *for all* i, j, k*, then the* $\mathbf{P}^K$*-distribution of* $\langle \mathfrak{X}(A_i)\rangle$ *equals the* $\mathbf{P}^L$*-distribution of* $\langle \mathfrak{X}(B_i)\rangle$*.*

Proof. When only finitely many $\lambda_k \neq 0$, this follows from our previous calculation. The general case follows from weak convergence of the processes corresponding to the partial sums, as in the paragraph following Lemma 3.1. □

This permits us to compare to discrete measures via [21, Lemma 16]:

Lemma 3.5. *Let* μ *be a Radon measure on a locally compact Polish space,* E*. Let* $\langle A_i\rangle$ *be pairwise disjoint Borel subsets of* E*. Let* $\phi_k \in L^2(E,\mu)$ *for* $k \geq 1$*. Then there exists a denumerable set* F*, pairwise disjoint subsets* $\langle B_i\rangle$ *of* F*, and* $v_k \in \ell^2(F)$ *such that* $(\phi_j,\ \phi_k) = (v_j,\ v_k)$ *and* $\big(\mathbf{1}_{A_i}\phi_j,\ \phi_k\big) = \big(\mathbf{1}_{B_i}v_j,\ v_k\big)$ *for all* i, j, k*.*

Proof. Without loss of generality, we may assume that $\bigcup_i A_i = E$. For each i, fix an orthonormal basis $\langle w_{i,j}\,;\ j < n_i\rangle$ for the subspace of $L^2(E,\mu)$ spanned by $\{\mathbf{1}_{A_i}\phi_j\}$. Here, $n_i \in \mathbb{N}\cup\{\infty\}$. Define $B_i := \{(i,j)\,;\ j < n_i\}$ and $F := \bigcup_i B_i$. Let T be the isometric isomorphism from the span of $\{w_{i,j}\,;\ i \geq 1,\ j < n_i\}$ to $\ell^2(F)$ that sends $w_{i,j}$ to $\mathbf{1}_{\{(i,j)\}}$. Defining $v_k := T(\phi_k)$ yields the desired vectors. □

3.7. Stochastic inequalities. We now show how the discrete models of Subsection 3.6 allow us to obtain the analogues of the stochastic inequalities known to hold for discrete determinantal probability measures.

For a Borel set $A \subseteq E$, let $\mathscr{F}(A)$ denote the σ-field on $\mathcal{N}(E)$ generated by the functions $\xi \mapsto \xi(B)$ for Borel $B \subseteq A$. We say that a function that is measurable with respect to $\mathscr{F}(A)$ is, more simply, measurable with respect to A. The obvious partial order on $\mathcal{N}(E)$ allows us to define what it means for a function $f\colon \mathcal{N}(E) \to \mathbb{R}$ to be ***increasing***. As in the discrete case, we say that $\mathbf{P}$ has ***negative associations*** if $\mathbf{E}[f_1 f_2] \leq \mathbf{E}[f_1]\mathbf{E}[f_2]$ for every pair f_1, f_2 of bounded increasing functions that are measurable with respect to complementary subsets of E. An event is increasing if its indicator is increasing. Then $\mathbf{P}$ has negative associations iff

$$\mathbf{P}(\mathcal{A}_1 \cap \mathcal{A}_2) \leq \mathbf{P}(\mathcal{A}_1)\mathbf{P}(\mathcal{A}_2) \tag{3.10}$$

for every pair $\mathcal{A}_1, \mathcal{A}_2$ of increasing events that are measurable with respect to complementary subsets of E.

We also say that $\mathbf{P}_1$ ***is stochastically dominated by*** $\mathbf{P}_2$ and write $\mathbf{P}_1 \preccurlyeq \mathbf{P}_2$ if $\mathbf{P}_1(\mathcal{A}) \leq \mathbf{P}_2(\mathcal{A})$ for every increasing event $\mathcal{A}$.

Call an event ***elementary increasing*** if it has the form $\{\xi\,;\ \xi(B) \geq k\}$, where B is a relatively compact Borel set and $k \in \mathbb{N}$. Write $\mathscr{U}(A)$ for the closure under finite unions and intersections of the collection of elementary increasing events with $B \subseteq A$; the notation $\mathscr{U}$ is chosen for "upwardly closed". Note that every event in $\mathscr{U}(A)$ is measurable with respect to some finite collection of functions $\xi \mapsto \xi(B_i)$ for pairwise *disjoint* relatively compact Borel $B_i \subseteq A$. Write $\overline{\mathscr{U}(A)}$ for the closure of $\mathscr{U}(A)$ under monotone limits, i.e., under unions of increasing sequences and under intersections of decreasing sequences; these events are also increasing. This is the same as the closure of $\mathscr{U}(A)$ under countable unions and intersections.

Lemma 3.6. *Let* A *be a Borel subset of a locally compact Polish space,* E*. Then* $\overline{\mathscr{U}(A)}$ *is exactly the class of increasing Borel sets in* $\mathscr{F}(A)$*.*

We give a proof at the end of this subsection. First, we derive two consequences. A weaker version (negative correlations of elementary increasing events) of the initial one is due to [20].

Theorem 3.7. *Let μ be a Radon measure on a locally compact Polish space, E. Let K be a locally trace-class positive contraction on $L^2(E,\mu)$. Then $\mathbf{P}^K$ has negative associations.*

Proof. Let $A \subset E$ be Borel. Let $\mathcal{A}_1 \in \mathscr{U}(A)$ and $\mathcal{A}_2 \in \mathscr{U}(E \setminus A)$. Then $\mathcal{A}_1, \mathcal{A}_2 \in \mathscr{F}(B)$ for some compact B by definition of $\mathscr{U}(\cdot)$. We claim that (3.10) holds for $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathbf{P} = \mathbf{P}^{K_B}$, i.e., for $\mathbf{P} = \mathbf{P}^K$.

Now $\mathcal{A}_1$ is measurable with respect to a finite number of functions $\xi \mapsto \xi(B_i)$ for some disjoint $B_i \subseteq A \cap B (1 \le i \le n)$ and $\mathcal{A}_2$ is measurable with respect to a finite number of functions $\xi \mapsto \xi(C_i)$ for some disjoint $C_i \subseteq B \setminus A (1 \le i \le n)$. Thus, there are functions g_1 and g_2 such that $\mathbf{1}_{\mathcal{A}_1}(\xi) = g_1\big(\xi(B_1), \dots, \xi(B_n)\big)$ and $\mathbf{1}_{\mathcal{A}_2}(\xi) = g_2\big(\xi(C_1), \dots, \xi(C_n)\big)$. By Theorem 3.4 and Lemma 3.5, there is some discrete determinantal probability measure $\mathbf{P}^Q$ on some denumerable set F and pairwise disjoint sets $B_i', C_i' \subseteq F$ such that the joint $\mathbf{P}^{K_B}$-distribution of all $\mathfrak{X}(B_i)$ and $\mathfrak{X}(C_i)$ is equal to the joint $\mathbf{P}^Q$-distribution of all $\mathfrak{X}(B_i')$ and $\mathfrak{X}(C_i')$. Define the corresponding events $\mathcal{A}_i'$ by $\mathbf{1}_{\mathcal{A}_1'}(\xi) = g_1\big(\xi(B_1'), \dots, \xi(B_n')\big)$ and $\mathbf{1}_{\mathcal{A}_2'}(\xi) = g_2\big(\xi(C_1'), \dots, \xi(C_n')\big)$. Since $\mathcal{A}_i'$ depend on disjoint subsets of F, Theorem 2.10 gives that $\mathbf{P}^Q(\mathcal{A}_1' \cap \mathcal{A}_2') \le \mathbf{P}^Q(\mathcal{A}_1')\mathbf{P}^Q(\mathcal{A}_2')$. This is the same as (3.10) by Theorem 3.4.

The same (3.10) clearly then holds in the less restrictive setting $\mathcal{A}_i \in \overline{\mathscr{U}(A)}$ by taking monotone limits. Lemma 3.6 completes the proof. □

Theorem 3.8 (Theorem 3 of [21]). *Suppose that K_1 and K_2 are two locally trace-class positive contractions such that $K_1 \preceq K_2$. Then $\mathbf{P}^{K_1} \preccurlyeq \mathbf{P}^{K_2}$.*

Proof. It suffices to show that $\mathbf{P}^{K_1}(\mathcal{A}) \le \mathbf{P}^{K_2}(\mathcal{A})$ for every $\mathcal{A} \in \mathscr{U}(E)$. Again, it suffices to assume that K_i are trace class. Lemma 3.5 applied to all eigenfunctions of K_1 and K_2 yields a denumerable F and two positive contractions K_i' on $\ell^2(F)$, together with an event $\mathcal{A}'$, such that $\mathbf{P}^{K_i'}(\mathcal{A}') = \mathbf{P}^{K_i}(\mathcal{A})$ for $i = 1, 2$. Furthermore, by construction, every function in $\ell^2(F)$ is the image of a function in $L^2(E)$ under the isometric isomorphism T used to prove Lemma 3.5, whence $K_1' \preceq K_2'$. Therefore Theorem 2.9 yields $\mathbf{P}^{K_1'}(\mathcal{A}') \le \mathbf{P}^{K_2'}(\mathcal{A}')$, as desired. □

Again, it would be very interesting to have a natural monotone coupling of $\mathbf{P}^{K_1}$ with $\mathbf{P}^{K_2}$. For some examples where this would be desirable, see Subsection 3.8.

Lemma 3.6 will follow from this folklore variant of a theorem of Dyck [16]:

Theorem 3.9. *Let X be a Polish space on which $\le$ is a partial ordering that is closed in $X \times X$. Let $\mathscr{U}$ be a collection of open increasing sets that generates the Borel subsets of X. Let $\mathscr{U}^*$ be the closure of $\mathscr{U}$ under countable intersections and countable unions. Suppose that for all $x, y \in X$, either $x \le y$ or there is $U \in \mathscr{U}$ and an open set $V \subset X$ such that $x \in U$, $y \in V$, and $U \cap V = \emptyset$. Then $\mathscr{U}^*$ equals the class of increasing Borel sets.*

Proof. Obviously every set in $\mathscr{U}^*$ is Borel and increasing. To show the converse, we prove a variant of Lusin's separation theorem. Namely, we show that if $W_1 \subset X$ is increasing and analytic (with respect to the paving of closed sets, as usual) and if $W_2 \subset X$ is analytic with $W_1 \cap W_2 = \emptyset$, then there exists $U \in \mathscr{U}^*$ such that $W_1 \subseteq U$ and $U \cap W_2 = \emptyset$. Taking W_1 to be Borel and $W_2 := X \setminus W_1$ forces $U = W_1$ and gives the desired conclusion.

To prove this separation property, we first show a stronger conclusion in a special case: Suppose that $A_1, A_2 \subset X$ are compact such that A_1 is contained in an increasing set W_1 that is disjoint from A_2; then there exists an open $U \in \mathscr{U}^*$ and an open V such that $A_1 \subseteq U$, $A_2 \subseteq V$, and $U \cap V = \emptyset$. Indeed, since W_1 is increasing, for every $(x, y) \in A_1 \times A_2$, we do *not* have that $x \le y$, whence by hypothesis, there exist $U_{x,y} \in \mathscr{U}$ and an open $V_{x,y}$ with $x \in U_{x,y}$, $y \in V_{x,y}$, and $U_{x,y} \cap V_{x,y} = \emptyset$. Because A_2 is compact, for each $x \in A_1$, we may choose $y_1, \dots, y_n \in A_2$ such that $A_2 \subseteq V_x := \bigcup_{i=1}^n V_{x,y_i}$. Define $U_x := \bigcap_{i=1}^n U_{x,y_i}$. Then U_x is open, contains x, and is disjoint from V_x, whence compactness of A_1 ensures the existence of $x_1, \dots, x_m \in A_1$ with $A_1 \subseteq U := \bigcup_{j=1}^m U_{x_j} \in \mathscr{U}^*$. Then $V := \bigcap_{j=1}^m V_{x_j}$ is open, contains A_2, and is disjoint from U, as desired.

To prove the general case, let π_1 and π_2 be the two coordinate projections on $X^2 = X \times X$. Define $I(A) = I\big(\pi_1(A) \times \pi_2(A)\big)$ for $A \subseteq X^2$ to be 0 if there exists $U \in \mathscr{U}^*$ such that $\pi_1(A) \subseteq U$ and $U \cap \pi_2(A) = \emptyset$; and to be 1 otherwise.

We claim that I is a capacity in the sense of [29, (30.1)]. It is obvious that $I(A) \le I(B)$ if $A \subseteq B$ and it is simple to check that if $A_1 \subseteq A_2 \subseteq \cdots$, then $\lim_{n\to\infty} I(A_n) = I\big(\bigcup_n A_n\big)$. Suppose for the final property that A is compact and $I(A) = 0$; we must find an open $B \supseteq A$ for which $I(B) = 0$. There exists some $W_1 \in \mathscr{U}^*$ with $\pi_1(A) \subseteq W_1$ and $W_1 \cap \pi_2(A) = \emptyset$. Then the result of the second paragraph yields sets U and V that give $B := U \times V$ as desired.

Now let W_1 and W_2 be as in the first paragraph. If $A \subseteq W_1 \times W_2$ is compact, then setting $A_i := \pi_i(A)$ and applying the second paragraph shows that $I(A) = 0$. Thus, by the Choquet capacitability theorem [29, (30.13)], $I(W_1 \times W_2) = 0$. □

Proof of Lemma 3.6. Clearly every set in $\overline{\mathscr{U}(A)}$ is increasing and in $\mathscr{F}(A)$. For the converse, endow A with a metric so that it becomes locally compact Polish while preserving its class of relatively compact sets and its Borel σ-field: Choose a denumerable partition of A into relatively compact sets A_i and make each one compact and of diameter at most 1; make the distance between x and y be 1 if x and y belong to different A_i. Let $X := \mathcal{N}(A)$ with the vague topology and let $\mathscr{U}$ be the class of elementary increasing events defined with respect to (relatively compact) sets $B \subseteq A$ that are open for this new metric. Apply Theorem 3.9. Since $\mathscr{U}^* \subseteq \overline{\mathscr{U}(A)}$, the result follows. □

3.8. Example: Orthogonal polynomial ensembles. Natural examples of determinantal point processes arise from orthogonal polynomials with respect to a probability measure μ on $\mathbb{C}$. Assume that μ has infinite support and finite moments of all orders. Let K_n denote the orthogonal projection of $L^2(\mathbb{C}, \mu)$ onto the linear span Poly_n of the functions $\{1, z, z^2, \dots, z^{n-1}\}$. There exist unique (up to signum) polynomials ϕ_k of degree k such that for every n, $\langle \phi_k \,;\ 0 \le k < n\rangle$ is an orthonormal basis of Poly_n. By elementary row operations, we see that for variables $(z_1, \dots, z_n)$, the map $(z_1, \dots, z_n) \mapsto \det[\phi_i(z_j)]_{i,j\le n}$ is a Vandermonde polynomial up to a constant factor, whence

$$\det(K_n\restriction\{z_1, \dots, z_n\}) = \det[\phi_i(z_j)][\phi_i(z_j)]^* = c_n \prod_{1\le i<j\le n} |z_i - z_j|^2$$

for some constant c_n. Therefore, the density of $\mathbf{P}^{K_n}$ (with points randomly ordered) with respect to μ^n is given by $c_n/n!$ times the square of a Vandermonde determinant.

Classical examples include the following:

OPE1. If μ is Gaussian measure on $\mathbb{R}$, i.e., $d\mu(x) = (2\pi)^{-1/2} e^{-x^2/2}\, dx$, then ϕ_k are the Hermite polynomials, $c_n = \big(\prod_{j=1}^{n-1} j!\big)^{-1}$, and $\mathbf{P}^{K_n}$ is the law of the ***Gaussian unitary***

ensemble, which is the set of eigenvalues of $(\mathfrak{M}+\mathfrak{M}^*)/\sqrt{2}$, where $\mathfrak{M}$ is an $n\times n$ matrix whose entries are independent standard complex Gaussian. (A standard complex Gaussian random variable is the same as a standard Gaussian vector in $\mathbb{R}^2$ divided by $\sqrt{2}$ in order that the complex variance equal 1. Its density is $\pi^{-1}e^{-|z|^2}$ with respect to Lebesgue measure on $\mathbb{C}$.) This is due to Wigner; see [40].

OPE2. If μ is unit Lebesgue measure on the unit circle $\{z\,;\ |z|=1\}$, then $\phi_k(z)=z^k$, so $c_n=1$, and $\mathbf{P}^{K_n}$ is the law of the ***circular unitary ensemble***, which is the set of eigenvalues of a random matrix whose distribution is Haar measure on the set of $n\times n$ unitary matrices. This ensemble was introduced by Dyson, but the law of the eigenvalues is due to Weyl; see [27].

OPE3. If μ is standard Gaussian measure on $\mathbb{C}$, then $\phi_k(z)=z^k/\sqrt{k!}$, $c_n=\left(\prod_{j=1}^{n-1}j!\right)^{-1}$, and $\mathbf{P}^{K_n}$ is the law of the n***th (complex) Ginibre process***, which is the set of eigenvalues of an $n\times n$ matrix whose entries are independent standard complex Gaussian. This is due to Ginibre; see [27].

OPE4. If μ is unit Lebesgue measure on the unit disk $\mathbb{D}:=\{z\,;\ |z|<1\}$, then $\phi_k(z)=\sqrt{k+1}\,z^k$, so $c_n=n!$, and the limit of $\mathbf{P}^{K_n}$ is the law of the zero set of the random power series whose coefficients are independent standard complex Gaussian, which converges in the unit disk a.s. This is due to Peres and Virág [45].

OPE5. If μ has density $z\mapsto n\pi^{-1}\bigl(1+|z|^2\bigr)^{-n-1}$ with respect to Lebesgue measure on $\mathbb{C}$, then $\phi_k(z)=\sqrt{\binom{n-1}{k}}z^k$ for $k<n$, so $c_n=\prod_{j=1}^{n-1}\binom{n-1}{j}$, and $\mathbf{P}^{K_n}$ is the law of the n***th spherical ensemble***, which is the set of eigenvalues of $\mathfrak{M}_1^{-1}\mathfrak{M}_2$ when $\mathfrak{M}_i$ are independent $n\times n$ matrices whose entries are independent standard complex Gaussian. (Here, we are limited to Poly_n since the larger spaces do not lie in $L^2(\mu)$.) This is due to Krishnapur [31]; see [27]. The process was studied earlier by [13] and [18], but without observing the connection to eigenvalues. Inverting stereographic projection, we identify this process with one whose density with respect to Lebesgue measure on the unit sphere in $\mathbb{R}^3$ is proportional to $\prod_{1\le i<j\le n}\|\mathbf{v}_i-\mathbf{v}_j\|^2$.

For additional information on such processes, see [50, 23, 47, 17]. For an extension to complex manifolds, see [3, 4, 5].

By Theorem 3.8, the processes $\mathbf{P}^{K_n}$ stochastically increase in n for each of the examples above except the last. It would be interesting to see natural monotone couplings. Perhaps the last example also increases stochastically in n.

The ***Ginibre process*** is the limit of the nth Ginibre processes as $n\to\infty$; it has the kernel $e^{z\bar{w}}$ with respect to standard Gaussian measure on $\mathbb{C}$. This process is invariant under all isometries of $\mathbb{C}$. For each of the plane, sphere, and hyperbolic disk, there is only a 1-parameter family of determinantal point processes having a kernel $K(z,w)$ that is holomorphic in z and in $\bar{w}$ and whose law is isometry invariant [31, Theorem 3.0.5]. For the sphere, that family has already been given above; the parameter is a positive integer. For the other two families, the parameter is a positive real number, α. In the case of the plane, the processes are related simply by homotheties, $M_\alpha\colon z\mapsto z/\alpha$. The push-forward of the Ginibre process with respect to $M_{\sqrt{\alpha}}$ has kernel $e^{\alpha z\bar{w}}$ with respect to the measure $\alpha\pi^{-1}e^{-\alpha|z|^2}d\mu(z)$, where μ is Lebesgue measure on $\mathbb{C}$. Do these processes increase stochastically in α, like Poisson processes do? In the hyperbolic disk, the processes have kernel $\alpha(1-z\bar{w})^{-\alpha-1}$ with respect to the measure $\pi^{-1}(1-|z|^2)^{\alpha-1}d\mu(z)$, where μ is

Lebesgue measure on $\mathbb{D}$. (We fix a branch of $(1-z)^{-\alpha-1}$ for $z \in \mathbb{D}$.) These give orthogonal projections onto the generalized Bergman spaces. The case $\alpha = 1$ is that of the limiting OPE4 above. Do these processes stochastically increase in α?

4. Completeness

Recall that when H is a finite-dimensional subspace of $\ell^2(E)$, the measure $\mathbf{P}^H$ is supported by those subsets $B \subseteq E$ that project to a basis of H under P_H. Similarly, when K is the kernel of a finite-rank orthogonal projection onto $H \subset L^2(E,\mu)$, define the functions $K_x := K(\cdot, x) = \sum_{k\geq 1} \overline{\phi_k(x)}\phi_k \in H$. Then the measure $\mathbf{P}^K$ is supported by those ξ such that $\langle K_x\,;\ x \in \xi\rangle$ is a basis of H, since $K(x,y) = (K_y,\, K_x)$. Here, $x \in \xi$ means that $\xi(\{x\}) = 1$.

The question of extending this to infinite-dimensional H turns out to be very interesting. A basis of a finite-dimensional vector space is a minimal spanning set. Although $P_H\mathfrak{B}$ is $\mathbf{P}^H$-a.s. linearly independent, minimality does not hold in general, even for the wired spanning forest of a tree, as shown by the examples in [24]. See also Corollary 4.5. However, the other half of being a basis does hold in the discrete case and is open in the continuous case.

4.1. Discrete completeness. Let $[V]$ be the closed linear span of $V \subseteq \ell^2(E)$.

Theorem 4.1 ([33]). *For every $H \leq \ell^2(E)$, we have $[P_H\mathfrak{B}] = H$ $\mathbf{P}^H$-a.s.*

We give an application of Theorem 4.1 for $E = \mathbb{Z}$, but it has an analogous statement for every countable abelian group. Let $\mathbb{T} := \mathbb{R}/\mathbb{Z}$ be the unit circle equipped with unit Lebesgue measure. For a measurable function $f\colon \mathbb{T} \to \mathbb{C}$ and $n \in \mathbb{Z}$, the ***Fourier coefficient*** of f at n is $\widehat{f}(n) := \int_{\mathbb{T}} f(t)e^{-2\pi int}\,dt$. Let $\widehat{f}{\restriction}S$ denote the restriction of $\widehat{f}$ to S. If $A \subseteq \mathbb{T}$ is measurable, we say $S \subseteq \mathbb{Z}$ is ***complete for*** A if the set $\{f\mathbf{1}_A\,;\ f \in L^2(\mathbb{T}),\ \widehat{f}{\restriction}(\mathbb{Z}\setminus S) \equiv 0\}$ is dense in $L^2(A)$, where we identify $L^2(A)$ with the set of functions in $L^2(\mathbb{T})$ that vanish outside A. The case where A is an interval is quite classical; see [46] for a review. A crucial role in that case is played by the following notion of density of S.

Definition 4.2. For an interval $[a,b] \subset \mathbb{R}\setminus\{0\}$, define its ***aspect***

$$\alpha\big([a,b]\big) := \max\big\{|a|,|b|\big\}/\min\big\{|a|,|b|\big\}\,.$$

For a discrete $S \subseteq \mathbb{R}$, the ***Beurling-Malliavin density*** of S, denoted $\mathsf{BM}(S)$, is the supremum of those $D \geq 0$ for which there exist disjoint nonempty intervals $I_n \subset \mathbb{R}\setminus\{0\}$ with $|S \cap I_n| \geq D|I_n|$ for all n and $\sum_{n\geq 1}[\alpha(I_n)-1]^2 = \infty$.

Corollary 4.3 ([33]). *Let $A \subset \mathbb{T}$ be Lebesgue measurable with measure $|A|$. Then there is a set of Beurling-Malliavin density $|A|$ in $\mathbb{Z}$ that is complete for A. Indeed, let $\mathbf{P}^A$ be the determinantal probability measure on $2^{\mathbb{Z}}$ corresponding to the Toeplitz matrix $(j,k) \mapsto \widehat{\mathbf{1}_A}(k-j)$. Then $\mathbf{P}^A$-a.e. $S \subset \mathbb{Z}$ is complete for A and has $\mathsf{BM}(S) = |A|$.*

When A is an interval, the celebrated theorem of Beurling and Malliavin [6] says that if S is complete for A, then $\mathsf{BM}(S) \geq |A|$, and that if $\mathsf{BM}(S) > |A|$, then S is complete for

A. (This holds for S that are not necessarily sets of integers, but we are concerned in this subsection only with $S \subseteq \mathbb{Z}$.)

Corollary 4.3 can be compared (take $\mathbb{T} \setminus A$ and $\mathbb{Z} \setminus S$) to a theorem of Bourgain and Tzafriri [9], according to which there is a set $S \subset \mathbb{Z}$ of (Schnirelman) density at least $2^{-8}|A|$ such that if $f \in L^2(\mathbb{T})$ and $\widehat{f}$ vanishes off S, then

$$|A|^{-1} \int_A |f(t)|^2 \, dt \geq 2^{-16} \|f\|_2^2 \,.$$

It would be interesting to find a quantitative strengthening of Corollary 4.3 that would encompass this theorem of [9].

The following theorem is equivalent to Theorem 4.1 by duality:

Theorem 4.4 ([33]). *For every $H \leq \ell^2(E)$, we have $\overline{P_{[\mathfrak{B}]} H} = [\mathfrak{B}]$ $\mathbf{P}^H$-a.s.*

As an example, consider the wired spanning forest of a graph, G. Here, $H := \bigstar(G)$. In this case, $H_B := \overline{P_{[B]} \bigstar(G)} = \bigstar(B)$ for $B \subseteq E$. Thus, the conclusion of Theorem 4.4 is that $\mathbf{P}^{H_\mathfrak{F}}$, which equals $\mathsf{WSF}_\mathfrak{F}$, is concentrated on the singleton $\{\mathfrak{F}\}$ for WSF_G-a.e. $\mathfrak{F}$. This was a conjecture of [2], established by [42].

Corollary 4.5. *For every $H \leq \ell^2(E)$, $\mathbf{P}^H$-a.s. the maps $P_H \colon [\mathfrak{B}] \to H$ and $P_{[\mathfrak{B}]} \colon H \to [\mathfrak{B}]$ are injective with dense image.*

Proof. Both statements are equivalent to $[\mathfrak{B}] \cap H^\perp = \{0\} = H \cap \mathfrak{B}^\perp$, and these are the contents of Theorems 4.1 and 4.4. □

4.2. Continuous completeness. If K is a locally trace-class orthogonal projection onto H, then for $h \in H$, we have

$$h(x) = (Kh)(x) = \int_E K(x,y) h(y) \, d\mu(y) = \int_E h(y) \overline{K(y,x)} \, d\mu(y) = \big(h, \ K_x\big) \,.$$

In other words, K is a reproducing kernel for H. A subset S of H is called ***complete for*** H if the closed linear span of S equals H; equivalently, the only element of H that is orthogonal to S is 0.

An analogue of Theorem 4.1 was conjectured by Lyons and Peres in 2010:

Conjecture 4.6. *If K is a locally trace-class orthogonal projection onto H, then for $\mathbf{P}^K$-a.e. $\mathfrak{X}$, $[\{K_x \,;\ x \in \mathfrak{X}\}] = H$, i.e., if $h \in H$ and $h \restriction \mathfrak{X} = 0$, then $h \equiv 0$.*

Just as in the discrete case, this appears to be on the critical border for many special instances, as we illustrate for several processes where $E = \mathbb{C}$:

1. Let μ be Lebesgue measure on $\mathbb{R}$ and $K(x,y) := \sin \pi(x-y) / \big(\pi(x-y)\big)$, the ***sine-kernel process***. Denote the Fourier transform on $\mathbb{R}$ by $\widehat{f}(t) := \int_{\mathbb{R}} f(x) e^{-2\pi i t x} \, dx$ for $f \in L^1(\mathbb{R})$, and, by isometric extension, for $f \in L^2(\mathbb{R})$. Write $I := \mathbf{1}_{[-1/2,1/2]}$. Since $K(x,0) = \widehat{I}(x)$, we have $(Kf)(x) = \big(f * \widehat{I}\big)(x) = \widehat{\check{f} I}(x)$, where $\check{f}$ is the inverse Fourier transform of f. Therefore, the induced operator K arises from the orthogonal projection onto the Paley-Wiener space $\{f \in L^2(\mathbb{R}, \mu) \,;\ \check{f}(t) = 0 \text{ if } |t| > 1/2\}$. The sine-kernel process arises frequently; e.g., it is various scaling limits of the

nth Gaussian unitary ensemble "in the bulk" as $n \to \infty$. (A related scaling limit of the GUE is Wigner's semicircle distribution.) We may more easily interpret Conjecture 4.6 for Fourier transforms of functions in $L^2[-1/2, 1/2]$: It says that for $\mathbf{P}^K$-a.e. $\mathfrak{X}$, the only $h \in L^2[-1/2, 1/2]$ such that $\widehat{h}\restriction\mathfrak{X} = 0$ is $h \equiv 0$. Although the Beurling-Malliavin theorem applies, no information can be deduced because $\mathsf{BM}(\mathfrak{X}) = 1$ a.s. However, Ghosh [20] has proved this case.

2. Let μ be standard Gaussian measure on $\mathbb{C}$ and $K(z, w) := e^{z\bar{w}}$. This is the Ginibre process. It corresponds to orthogonal projection onto the ***Bargmann-Fock space*** $B^2(\mathbb{C})$ consisting of the entire functions that lie in $L^2(\mathbb{C}, \mu)$; this is the space of power series $\sum_{n\geq 0} a_n z^n$ such that $\sum_n n!|a_n|^2 < \infty$. Completeness of a set of elements $\{e^{\lambda z}\,;\ \lambda \in \Lambda\} \subset B^2(\mathbb{C})$ in $B^2(\mathbb{C})$ is equivalent to completeness in $L^2(\mathbb{R})$ (with Lebesgue measure) of the Gabor system of windowed complex exponentials

$$\left\{t \mapsto \exp\left[-i \operatorname{Im} \lambda t - (t - \operatorname{Re} \lambda)^2\right]\,;\ \lambda \in \sqrt{2}\Lambda\right\},$$

which is used in time-frequency analysis of non-band-limited signals. The equivalence is proved using the Bargmann transform

$$f \mapsto \left(z \mapsto \pi^{-1/4} \int_{\mathbb{R}} f(t) \exp\left[\sqrt{2}tz - \frac{z^2}{2} - \frac{t^2}{2}\right] dt\right),$$

which is an isometry from $L^2(\mathbb{R})$ to $B^2(\mathbb{C})$. That the critical density is 1 was shown in various senses going back to von Neumann; see [14]. This case has also been proved by Ghosh [20].

3. Let μ be unit Lebesgue measure on the unit disk $\mathbb{D} := \{z\,;\ |z| < 1\}$ and $K(z, w) := \left(1 - z\bar{w}\right)^{-2}$. This process is the limiting OPE4 in Subsection 3.8. It corresponds to orthogonal projection onto the ***Bergman space*** $A^2(\mathbb{D})$ consisting of the analytic functions that lie in $L^2(\mathbb{D}, \mu)$. What is known about the zero sets of functions in the Bergman space [15] is insufficient to settle Conjecture 4.6 in this case and it remains open.

The two instances above that have been proved by Ghosh [20] follow from his more general result that Conjecture 4.6 holds whenever μ is continuous and $\mathbf{P}^K$ is ***rigid***, which means that $\mathfrak{X}(B)$ is measurable with respect to the $\mathbf{P}^K$-completion of $\mathscr{F}(E \setminus B)$ for every ball $B \subset E$. The limiting process OPE4 is not rigid [25]. Ghosh and Krishnapur (personal communication, 2014) have shown that $\mathbf{P}^K$ is rigid only if K is an orthogonal projection. It is not sufficient that K be a projection, as the example of the Bergman space shows. A necessary and sufficient condition to be rigid is not known.

Let K be a locally trace-class orthogonal projection onto $H \leq L^2(E, \mu)$. For a function f, write f_K for the function $f(x)/\sqrt{K(x, x)}$. Let $\mathfrak{X} \sim \mathbf{P}^K$. Clearly $f_K\restriction\mathfrak{X} \in \ell^2(\mathfrak{X})$ for a.e. $\mathfrak{X}$. Also, for $h \in H$, the function h_K is bounded. A conjecture analogous to Corollary 4.5 is that $\mathfrak{X}$ is a sort of set of interpolation for H in the sense that given any countable dense set $H_0 \subset H$, for a.e. $\mathfrak{X}$, the set $\{h_K\restriction\mathfrak{X}\,;\ h \in H_0\}$ is dense in $\ell^2(\mathfrak{X})$.

One may also ask about completeness for appropriate Poisson point processes.

5. Discrete invariance

Suppose Γ is a group that acts on E and that K is Γ-invariant, i.e., $K(\gamma x, \gamma y) = K(x, y)$ for all $\gamma \in \Gamma$, $x \in E$, and $y \in E$. (This is equivalent to the operator K being Γ-equivariant.) Then the probability measure $\mathbf{P}^K$ is Γ-invariant. This contact with ergodic theory and other areas of mathematics suggests many interesting questions. Lack of space prevents us from considering more than just a few aspects of the case where E is discrete and from giving all definitions.

5.1. Integer lattices. Let $E := \Gamma := \mathbb{Z}^d$. In this case, K is invariant iff $K(m, n) = \widehat{f}(n - m)$ for some $f\colon \mathbb{T}^d \to [0, 1]$, where $\widehat{f}(n) := \int_{\mathbb{T}^d} f(t) e^{-2\pi i n \cdot t}\, dt$. We write $\mathbf{P}^f$ in place of $\mathbf{P}^K$. Some results and questions from [37] follow.

Theorem 5.1. *For all f, the process $\mathbf{P}^f$ is isomorphic to a Bernoulli process.*

This was shown in dimension 1 by [49] for those f such that $\sum_{n \geq 1} n |\widehat{f}(n)|^2 < \infty$ by showing that those $\mathbf{P}^f$ are weak Bernoulli (WB), also called "β-mixing" and "absolutely regular". Despite its name, it is known that WB is strictly stronger than Bernoullicity. The precise class of f for which $\mathbf{P}^f$ is WB is not known.

As usual, the ***geometric mean*** of a nonnegative function f is $\mathsf{GM}(f) := \exp \int \log f$.

Theorem 5.2. *For all f, the process $\mathbf{P}^f$ stochastically dominates product measure $\mathbf{P}^{\mathsf{GM}(f)}$ and is stochastically dominated by product measure $\mathbf{P}^{1-\mathsf{GM}(\mathbf{1}-f)}$. These bounds are optimal.*

We conjecture that (Kolmogorov-Sinai) entropy is concave, as would follow from Conjecture 2.6.

Conjecture 5.3. *For all f and g, we have $H\big(\mathbf{P}^{(f+g)/2}\big) \geq \big(H(\mathbf{P}^f) + H(\mathbf{P}^g)\big)/2$.*

Question 5.4. *Let $f\colon \mathbb{T} \to [0, 1]$ be a trigonometric polynomial of degree m. Then $\mathbf{P}^f$ is m-dependent, as are all $(m+1)$-block factors of independent processes. Is $\mathbf{P}^f$ an $(m+1)$-block factor of an i.i.d. process? This is known when $m = 1$ [10].*

5.2. Sofic groups. Let Γ be a sofic group, a class of groups that includes all finitely generated amenable groups and all finitely generated residually amenable groups. No finitely generated group is known not to be sofic. Let E be Γ or, more generally, a set acted on by Γ with finitely many orbits, such as the edges of a Cayley graph of Γ. The following theorems are from [38].

Theorem 5.5. *For every Γ-equivariant positive contraction Q on $\ell^2(E)$, the process $\mathbf{P}^Q$ is a $\bar{d}$-limit of finitely dependent (invariant) processes. If Γ is amenable and $E = \Gamma$, then $\mathbf{P}^Q$ is isomorphic to a Bernoulli process.*

Even if $\mathbf{P}^1$ and $\mathbf{P}^2$ are Γ-invariant probability measures on 2^Γ with $\mathbf{P}^1 \preccurlyeq \mathbf{P}^2$, there need not be a Γ-invariant monotone coupling of $\mathbf{P}^1$ and $\mathbf{P}^2$ [41]. The proof of the preceding theorem depends on the next one:

Theorem 5.6. *If Q_1 and Q_2 are two Γ-equivariant positive contractions on $\ell^2(E)$ with $Q_1 \preceq Q_2$, then there exists a Γ-invariant monotone coupling of $\mathbf{P}^{Q_1}$ and $\mathbf{P}^{Q_2}$.*

The proof of Theorem 5.5 also uses the inequality

$$\bar{d}\big(\mathbf{P}^Q, \mathbf{P}^{Q'}\big) \le 6 \cdot 3^{2/3} \|Q - Q'\|_1^{1/3}$$

for equivariant positive contractions, Q and Q', where $\|T\|_1 := \operatorname{tr}(T^*T)^{1/2}$ is the Schatten 1-norm. When Q and Q' commute, one can improve this bound to

$$\bar{d}(\mathbf{P}^Q, \mathbf{P}^{Q'}) \le \|Q - Q'\|_1 .$$

We do not know whether this inequality always holds.

Write $\mathsf{FK}(Q) := \exp \operatorname{tr} \log |Q|$ for the Fuglede-Kadison determinant of Q when Q is a Γ-equivariant operator. The following would extend Theorem 5.2. It is open even for finite groups.

Conjecture 5.7. *For all Γ-equivariant positive contractions Q on $\ell^2(\Gamma)$, the process $\mathbf{P}^Q$ stochastically dominates product measure $\mathbf{P}^{\mathsf{FK}(Q)I}$ and is stochastically dominated by product measure $\mathbf{P}^{I-\mathsf{FK}(I-Q)I}$, and these bounds are optimal.*

5.3. Isoperimetry, cost, and ℓ^2-Betti numbers. It turns out that the expected degree of a vertex in the free uniform spanning forest of a Cayley graph depends only on the group, via its first ℓ^2-Betti number, $\beta_1(\Gamma)$, and not on the generating set used to define the Cayley graph [34]:

Theorem 5.8. *In every Cayley graph G of a group Γ, we have*

$$\mathbf{E}_{\mathsf{FSF}(G)}[\deg_{\mathfrak{F}}(o)] = 2\beta_1(\Gamma) + 2 .$$

This is proved using the representation of FSF as a determinantal probability measure. It can be used to give a uniform bound on expansion constants [36]:

Theorem 5.9. *For every finite symmetric generating set S of a group Γ, we have $|SA \setminus A| > 2\beta_1(\Gamma)|A|$ for all finite non-empty $A \subset \Gamma$.*

There are extensions of these results to higher-dimensional CW-complexes and higher ℓ^2-Betti numbers [34].

In unpublished work with D. Gaboriau [35], we have shown the following:

Theorem 5.10. *Let G be a Cayley graph of a finitely generated group Γ and $\epsilon > 0$. Then there exists a Γ-invariant finitely dependent determinantal probability measure $\mathbf{P}^Q$ on $\{0,1\}^{\mathsf{E}(G)}$ that stochastically dominates FSF_G and such that*

$$\mathbf{E}^Q\big[\deg_{\mathfrak{S}}(o)\big] \le \mathbf{E}_{\mathsf{FSF}}\big[\deg_{\mathfrak{F}}(o)\big] + \epsilon .$$

In addition, if Γ is sofic, then $\bar{d}(\mathbf{P}^Q, \mathsf{FSF}) \le \epsilon$.

If it could be shown that $\mathbf{P}^Q$, or indeed every invariant finitely dependent probability measure that dominates FSF, yields a connected subgraph a.s., then it would follow that $\beta_1(\Gamma) + 1$ is equal to the cost of Γ, a major open problem of [19].

Acknowledgments. Partially supported by NSF grant DMS-1007244. I am grateful to Alekos Kechris for informing me of Theorem 3.9; the proof given seems to be due to Alain Louveau. I thank Norm Levenberg for references.

References

[1] Bapat, R.B., *Mixed discriminants and spanning trees*, Sankhyā Ser. A **54** (1992), no. Special Issue, 49–55, Combinatorial mathematics and applications (Calcutta, 1988).

[2] Benjamini, I., Lyons, R., Peres, Y., and Schramm, O., *Uniform spanning forests*, Ann. Probab. **29** (2001), 1–65.

[3] Berman, R.J., *Determinantal point processes and fermions on complex manifolds: Bulk universality*, Preprint, http://www.arxiv.org/abs/0811.3341, (2008).

[4] ———, *Determinantal point processes and fermions on complex manifolds: large deviations and bosonization*, Preprint, http://www.arxiv.org/abs/0812.4224, (2008).

[5] ———, *Sharp asymptotics for Toeplitz determinants*, Int. Math. Res. Not. IMRN **2012** (2012), 22, 5031–5062.

[6] Beurling, A. and Malliavin, P., *On the closure of characters and the zeros of entire functions*, Acta Math. **118** (1967), 79–93.

[7] Borcea, J., Brändén, P., and Liggett, T.M., *Negative dependence and the geometry of polynomials*, J. Amer. Math. Soc. **22** (2009), 521–567.

[8] Borodin, A., *Determinantal point processes*, The Oxford Handbook of Random Matrix Theory, pp. 231–249, Oxford Univ. Press, Oxford, 2011.

[9] Bourgain, J. and Tzafriri, L., *Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis*, Israel J. Math. **57** (1987), 2, 137–224.

[10] Broman, E., *One-dependent trigonometric determinantal processes are two-block-factors*, Ann. Probab. **33** (2005), 2, 601–609.

[11] Brooks, R.L., Smith, C.A.B., Stone, A.H., and Tutte, W.T., *The dissection of rectangles into squares*, Duke Math. J. **7** (1940), 312–340.

[12] Burton, R.M. and Pemantle, R., *Local characteristics, entropy and limit theorems for spanning trees and domino tilings via transfer-impedances*, Ann. Probab. **21** (1993), 1329–1371.

[13] Caillol, J.M., *Exact results for a two-dimensional one-component plasma on a sphere*, J. Physique - LETTRES **42** (1981), L-245–L-247.

[14] Chistyakov, G., Lyubarskii, Yu. and Pastur, L., *On completeness of random exponentials in the Bargmann-Fock space*, J. Math. Phys. **42** (2001), 8, 3754-3768.

[15] Duren, P. and Schuster, A., *Bergman Spaces*, Mathematical Surveys and Monographs, **100**. American Mathematical Society, Providence, RI, 2004.

[16] Dyck, S., *Some applications of positive formulas in descriptive set theory and logic*, Ann. Pure Appl. Logic **46** (1990), 2, 95–146.

[17] Forrester, P.J., *Log-Gases and Random Matrices*, London Mathematical Society Monographs Series, **34**. Princeton University Press, Princeton, NJ, 2010.

[18] Forrester, P.J., Jancovici, B., and Madore, J., *The two-dimensional Coulomb gas on a sphere: exact results*, J. Statist. Phys. **69** (1992), 1–2, 179–192.

[19] Gaboriau, D., *Invariants ℓ^2 de relations d'équivalence et de groupes*, Publ. Math. Inst. Hautes Études Sci. **95** (2002), 93–150.

[20] Ghosh, S., *Determinantal processes and completeness of random exponentials: the critical case*, Preprint, http://www.arxiv.org/abs/1211.2435, (2012).

[21] Goldman, A., *The Palm measure and the Voronoi tessellation for the Ginibre process*, Ann. Appl. Probab. **20** (2010), 1, 90–128.

[22] Häggström, O., *Random-cluster measures and uniform spanning trees*, Stochastic Process. Appl. **59** (1995), 267–275.

[23] Hardy, A., *Average characteristic polynomials of determinantal point processes*, Preprint, http://www.arxiv.org/abs/1211.6564, (2012).

[24] Heicklen, D. and Lyons, R., *Change intolerance in spanning forests*, J. Theoret. Probab. **16** (2003), 47–58.

[25] Holroyd, A.E. and Soo, T., *Insertion and deletion tolerance of point processes*, Electron. J. Probab. **18** (2013), 74, 24 pp.

[26] Hough, J.B., Krishnapur, M., Peres, Y., and Virág, B., *Determinantal processes and independence*, Probab. Surv. **3** (2006), 206–229.

[27] ———, *Zeros of Gaussian Analytic Functions and Determinantal Point Processes*, University Lecture Series, **51**. American Mathematical Society, Providence, RI, 2009.

[28] Johansson, K., *Random matrices and determinantal processes*, Mathematical Statistical Physics, 1–55, Elsevier B. V., Amsterdam, 2006.

[29] Kechris, A.S., *Classical Descriptive Set Theory*, Graduate Texts in Mathematics, **156**, Springer-Verlag, New York, 1995.

[30] Kirchhoff, G., *Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird*, Ann. Phys. und Chem. **72** (1847), 497–508.

[31] Krishnapur, M., *Zeros of Random Analytic Functions*, Ph.D. Thesis, U.C. Berkeley, 2006. http://www.arxiv.org/abs/math/0607504.

[32] Kulesza, A. and Taskar, B., *Determinantal point processes for machine learning*, Foundations and Trends in Machine Learning, **5** (2012), 2–3, 123–286. DOI:10.1561/2200000044.

[33] Lyons, R., *Determinantal probability measures*, Publ. Math. Inst. Hautes Études Sci. **98** (2003), 167–212. Errata, http://pages.iu.edu/~rdlyons/errata/bases.pdf.

[34] ———, *Random complexes and ℓ^2-Betti numbers*, J. Topology Anal. **1** (2009), 2, 153–175.

[35] Lyons, R. and Gaboriau, D., *An approach to the cost vs. ℓ^2-Betti-numbers problem*, In preparation.

[36] Lyons, R., Pichot, M., and Vassout, S., *Uniform non-amenability, cost, and the first ℓ^2-Betti number*, Groups Geom. Dyn. **2** (2008), 4, 595–617.

[37] Lyons, R. and Steif, J.E., *Stationary determinantal processes: Phase multiplicity, Bernoullicity, entropy, and domination*, Duke Math. J., **120** (2003), 3, 515–575.

[38] Lyons, R. and Thom, A., *Invariant coupling of determinantal measures on sofic groups*, Ergodic Theory Dynam. Systems, to appear. Preprint, http://www.arxiv.org/abs/1402.0969, (2014).

[39] Macchi, O., *The coincidence approach to stochastic point processes*, Advances in Appl. Probability **7** (1975), 83–122.

[40] Mehta, M., *Random Matrices*. Third edition. Pure and Applied Mathematics (Amsterdam), **142**. Elsevier/Academic Press, Amsterdam, 2004.

[41] Mester, P., *Invariant monotone coupling need not exist*, Ann. Probab. **41** (2013), 3A, 1180–1190.

[42] Morris, B., *The components of the wired spanning forest are recurrent*, Probab. Theory Related Fields **125** (2003), 259–265.

[43] Paulsen, V., *Completely Bounded Maps and Operator Algebras*, Cambridge Studies in Advanced Mathematics, **78**. Cambridge University Press, Cambridge, 2002.

[44] Pemantle, R., *Choosing a spanning tree for the integer lattice uniformly*, Ann. Probab. **19** (1991), 1559–1574.

[45] Peres, Y. and Virág, B., *Zeros of the i.i.d. Gaussian power series: a conformally invariant determinantal process*, Acta Math. **194** (2005), 1, 1–35.

[46] Redheffer, R., *Completeness of sets of complex exponentials*, Advances in Math. **24** (1977), 1–62.

[47] Rider, B. and Virág, B., *Complex determinantal processes and H^1 noise*, Electron. J. Probab. **12** (2007), 45, 1238–1257.

[48] Shirai, T. and Takahashi, Y., *Random point fields associated with certain Fredholm determinants. I. Fermion, Poisson and boson point processes*, J. Funct. Anal. **205** (2003), 414–463.

[49] ———, *Random point fields associated with certain Fredholm determinants II: fermion shifts and their ergodic and Gibbs properties*, Ann. Probab. **31** (2003), 1533–1564.

[50] Simon, B., *Weak convergence of CD kernels and applications*, Duke Math. J. **146** (2009), 2, 305–330.

[51] Soshnikov, A., *Determinantal random point fields*, Uspekhi Mat. Nauk **55** (2000), 107–160.

[52] ———, *Gaussian limit for determinantal random point fields*, Ann. Probab. **30** (2002), 1, 171–187.

[53] Strassen, V., *The existence of probability measures with given marginals*, Ann. Math. Statist. **36** (1965), 423–439.

Dept. of Math., Indiana University, 831 E. 3rd St., Bloomington, IN 47405-7106 USA
E-mail: rdlyons@indiana.edu

Rough paths, signatures and the modelling of functions on streams

Terry Lyons

Abstract. Rough path theory is focused on capturing and making precise the interactions between highly oscillatory and non-linear systems. The techniques draw particularly on the analysis of LC Young and the geometric algebra of KT Chen. The concepts and theorems, and the uniform estimates, have found widespread application; the first applications gave simplified proofs of basic questions from the large deviation theory and substantially extending Ito's theory of SDEs; the recent applications contribute to (Graham) automated recognition of Chinese handwriting and (Hairer) formulation of appropriate SPDEs to model randomly evolving interfaces. At the heart of the mathematics is the challenge of describing a smooth but potentially highly oscillatory and vector valued path x_t parsimoniously so as to effectively predict the response of a nonlinear system such as $dy_t = f(y_t)dx_t, y_0 = a$. The Signature is a homomorphism from the monoid of paths into the grouplike elements of a closed tensor algebra. It provides a graduated summary of the path x. Hambly and Lyons have shown that this non-commutative transform is faithful for paths of bounded variation up to appropriate null modifications. Among paths of bounded variation with given Signature there is always a unique shortest representative. These graduated summaries or features of a path are at the heart of the definition of a rough path; locally they remove the need to look at the fine structure of the path. Taylor's theorem explains how any smooth function can, locally, be expressed as a linear combination of certain special functions (monomials based at that point). Coordinate iterated integrals form a more subtle algebra of features that can describe a stream or path in an analogous way; they allow a definition of rough path and a natural linear "basis" for functions on streams that can be used for machine learning.

Mathematics Subject Classification (2010). 93C15, 68Q32, 60H10, 34F05, 60H15.

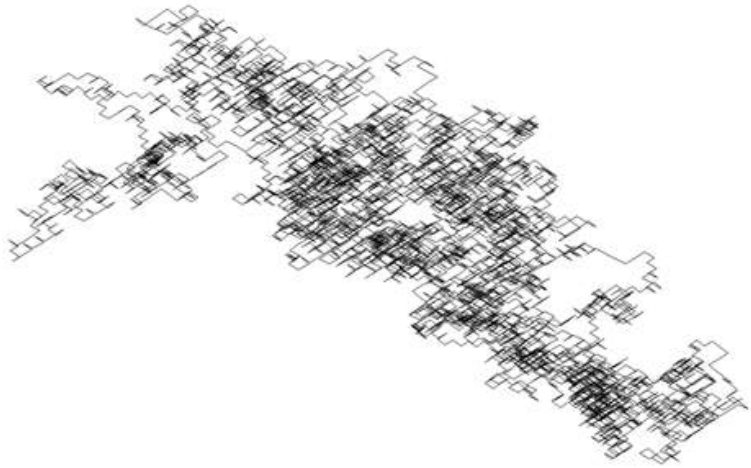
Keywords. Rough paths, regularity structures, machine learning, functional regression, numerical approximation of parabolic PDE, shuffle product, tensor algebra.

1. A path or a text?

The mathematical concept of a path embraces the notion of an evolving or time ordered sequence of events, parameterised by a continuous variable. Our mathematical study of these objects does not encourage us to think broadly about the truly enormous range of "paths" that occur. This talk will take an analyst's perspective, we do not expect to study a particular path but rather to find broad brush tools that allow us to study a wide variety of paths - ranging form very "pure" mathematical objects that capture holonomy to very concrete paths that describe financial data. Our goal will be to explain the progress we have made in the last 50 years or so in describing such paths effectively, and some of the consequences of these developments.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Let us start by noting that although most mathematicians would agree on a definition of a path, most have a rather stereotyped and limited imagination about the variety of paths that are "in the wild". One key observation is that in most cases we are interested in paths because they represent some evolution that interacts with and influences some wider system. Another is that in most paths, in standard presentations, the content and influence are locked into complex multidimensional oscillations.



The path in the figure is a piece of text. Each character in the text is encoded using ascii as a byte of 8 bits, each byte is represented as four letters of two bits, each two bit letter is represented by a line from the centre to one of the four corners of a square (for visial reasons the centre of this square is dispaced slightly to create a loop). The text can easily be represented in other ways, perhaps in different font or with each character as a bitmap. Each stream has broadly the same effect on a coarse scale although the detailed texture is perhaps a bit different.

2. Financial data or a semimartingale

One important source of sequential data comes from financial markets. An intrinsic feature of financial markets is that they are high dimensional but there is a strong notion of sequencing of events. Buying with future knowledge is forbidden. Much of the information relates to prices, and one of the radical successes of applied mathematics over the last 20-30 years came out of the approximation of price processes by simple stochastic differential equations and semimartingales and the use of Itô's calculus. However, modern markets are not represented by simple price processes. Most orders happen on exchanges, where there are numerous bids, offers, and less commonly, trades. Much activity in markets is concerned with market making and the provision of liquidity; decisions to post to the market are based closely on expectation of patterns of behaviour, and most decisions are somewhat distant from any view about fundamental value. If one is interested in alerting the trader who has a bug in his code, or understanding how to trade a large order without excessive charges then the semi-martingale model has a misplaced focus.

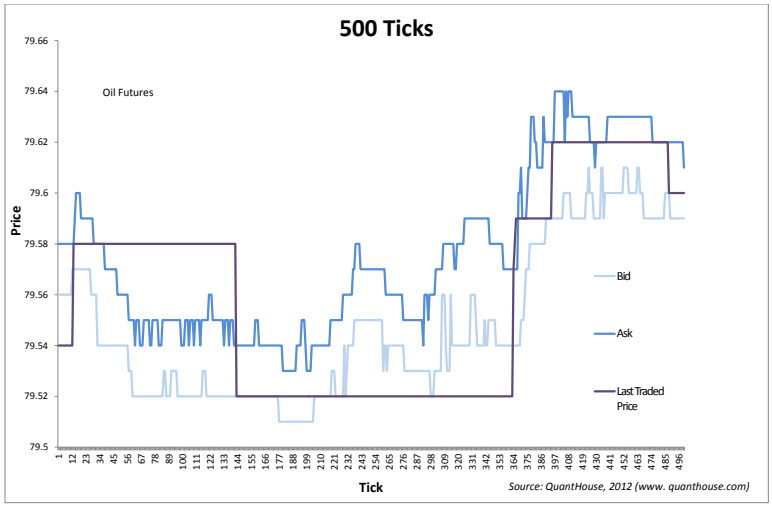


Figure 2.1. A snapshot of level one order book data

The data in the Figure 2.1 is a snapshot of the level one order book showing activity on a market for oil futures over 500 changes (roughly a 15 minute period). One can see the bid and offer prices changing, although trades happen (and so the last executed price changes) much less frequently. It is questionable whether a semi-martingale model for prices can capture this rich structure effectively.

3. Paths - simply everywhere - evolving systems

Informally, a stream is a map γ from a totally ordered set I to some state space, where we are interested in the effect (or transformation of state) this stream achieves. As we have noted the same stream of information can admit different representations with different fidelity. When the totally ordered set I is an interval and there are reasonable path properties (e.g. such as right continuity) we will call the stream a path. Nonetheless, many interesting streams are finite and discrete. There are canonical and informative ways to convert them [10] to continuous paths.

It is worth noting that, even at this abstract level, there are natural mathematical operations and invariances that are applied to a stream. One can reparameterise the speed at which one examines the stream and simultaneously the speed at which one looks at the effects. One can split a stream into two or more segments (a coproduct). One can sub-sample a stream. In general we will focus on those streams which are presented in a way where such sub-sampling degrades the information in the stream gradually. One can also merge or interleave discrete streams according to their time stamps if the totally ordered sets I, I'can be interleaved. All of these properties are inherited for the properties of totally ordered sets. If the target "effect" or state space is linear there is also the opportunity to translate and so concatenate streams or paths [15] and so get richer algebraic structures. One of the most

interesting and economically important questions one can ask about a stream is how to summarise (throw away irrelevant information) so as to succinctly capture its effects. We give a few examples in Table 3.1.

text	schoolchild	precis
sound	audio engineer	faithful perception
web page	search provider	interest for reader
web click history	advertiser	effective ad placement
Brownian path	numerical analysis	effective simulation
rough paths	analyst	RDEs

Table 3.1. Examples of contexts where streams are summarised while retaining their essence.

What is actually quite surprising is that there is a certain amount of useful work one can do on this problem that does not depend on the nature of the stream or path.

4. A simple model for an interacting system

We now focus on a very specific framework where the streams are maps from a real interval, that we will intuitively refer to as the time domain, into an a Banach space that we will refer to as the state space. We will work with continuous paths in continuous time but, as we mentioned, there are canonical ways to embed discrete tick style data into this framework using the Hoff process and in financial contexts this is important. There is also a more general theory dealing with paths with jumps [Williams, Simon].

4.1. Controlled differential equations. A path is a map γ from an interval $J = [J_-, J_+]$ into a Banach space E. The dimension of E may well be finite, but we allow for the possibility that it is not. It has bounded (p-)variation if

$$\begin{aligned} \sup_{\ldots u_i < u_{i+1} \ldots \in [J_-, J_+]} \sum_i \left\| \gamma_{u_{i+1}} - \gamma_u \right\| &< \infty \\ \sup_{\ldots u_i < u_{i+1} \ldots \in [J_-, J_+]} \sum_i \left\| \gamma_{u_{i+1}} - \gamma_u \right\|^p &< \infty \end{aligned}$$

where $p \geq 1$ In our context the path γ is controlling the system, and we are interested in its effect as measured by y and the interactions between γ and y. It would be possible to use the theory of rough paths to deal with the internal interactions of autonomous and "rough" systems, one specific example of deterministic McKean Vlasov type is [4].

Separately there needs to be a space F that carries the state of the system and a family of different ways to evolve. We represent the dynamics on F through the space $\Omega(F)$ of vector fields on F. Each vector field provides a different way for the state to evolve. We connect this potential to evolve the state in F to the control γ via a linear map

$$V : E \overset{linear}{\to} \Omega(F).$$

Immediately we can see the controlled differential equation

$$dy_t = V(y_t)\, d\gamma_t,\ y_{J_-} = a$$

$$\pi_J\left(y_{J_-}\right) \quad := \quad y_{J_+}$$

provides a precise framework allowing for the system y to respond to γ according to the dynamics V. We call such a system a controlled differential equation.

The model of a controlled differential equation is a good one. Many different types of object can be positioned to fit the definition. Apart from the more obvious applied examples, one can view a finite automata (in computer science sense) and the geometric concept of lifting a path along a connection as producing examples.

There are certain apparently trivial properties that controlled differential equations and the paths that control them have; none the less they are structurally essential so we mention them now.

Lemma 4.1 (Reparameterisation). *If $\tau : I \to J$ is an increasing homeomorphism, and if*

$$dy_t = V\left(y_t\right) d\gamma_t,\ y_{J_-} = a,$$

then the reparameterised control produces the reparameterised effect:

$$dy_{\tau(t)} = V\left(y_{\tau(t)}\right) d\gamma_{\tau(t)},\ y_{\tau(I_-)} = a.$$

Lemma 4.2 (Splitting). *Let π_J be the diffeomorphism capturing the transformational effect of $\gamma|_J$. Let $t \in J$. Then π_J can be recovered by composing the diffeomorpisms $\pi_{[J_-,t]}$, $\pi_{[t,J_+]}$ associated with splitting the interval J at t and considering the composing the effect of $\gamma|_{[J_-,t]}$ and $\gamma|_{[t,J_+]}$ separately:*

$$\pi_{[t,J_+]}\pi_{[J_-,t]} = \pi_J.$$

In this way we see that, assuming the vector fields were smooth enough to solve the differential equations uniquely and for all time, a controlled differential equation is a homomorphism from the monoid of paths with concatenation into the diffeomorphisms/transformations of the state space. By letting π act as an operator on functions we see that every choice of V defines a representation of the monoid of paths in E

Remark 4.3 (Subsampling). Although there is a good behaviour with respect to sub-sampling, which in effect captures and quantifies the numerical analysis of these equations, it is more subtle and we do not make it explicit here.

Remark 4.4. Fixing V, restricting γ to smooth paths on $[0, 1]$ and considering the solutions y with $y_0 = a$, generically the closure of the set of pairs (γ, y) in the uniform topology is NOT the graph of a map; $\gamma \to y$ is not closable and so is not well defined as a (even an unbounded and discontinuous) function in the space of continuous paths. Different approximations lead to different views as to what the solution should be.

4.2. Linear controlled differential equations. Where the control γ is fixed and smooth, the state space is linear, and all the vector fields are linear, then the space of responses y, as one varies the starting location a, is a linear space and $\pi_{[S,T]} : a = y_S \to y_T$ is a linear automorphism. This case is essentially Cartan's development of a path in a Lie Algebra into a path in the Lie Group starting at the identity. From our point of view it is a very important special case of our controlled differential equations; it reveals one of the key objects we want to discuss in this paper.

Suppose F is a Banach space, and A is a linear map $E \to Hom_{\mathbb{R}}(F,F)$ and that γ_t is a path in E. Consider the linear differential equation

$$dy_t = Ay_t d\gamma_t.$$

By iterating using Picard iteration one obtains

$$y_{J_+} = \left(\sum_{n=0}^{\infty} A^n \int\limits_{J_- \le u_1 \le \ldots \le u_n \le J_+} \cdots \int d\gamma_{u_1} \otimes \ldots \otimes d\gamma_{u_n} \right) y_0$$

The Signature of γ over the interval $J = [J_-, J_+]$

Definition 4.5. The Signature S of a bounded variation path (or more generally a weakly geometric p-rough path) γ over the interval $J = [J_-, J_+]$ is the tensor sequence

$$S(\gamma|_J) := \sum_{n=0}^{\infty} \int\limits_{u_1 \le \ldots \le u_n \in J^n} \cdots \int d\gamma_{u_1} \otimes \ldots \otimes d\gamma_{u_n} \in \bigoplus_{n=0}^{\infty} E^{\otimes n}$$

It is sometimes written $S(\gamma)_J$ or $S(\gamma)_{J_-,J_+}$.

Lemma 4.6. *The path $t \to S(\gamma)_{0,t}$ solves a linear differential equation controlled by γ.*

Proof. The equation is the universal non-commutative exponential:

$$\begin{aligned} dS_{0,t} &= S_{0,t} \otimes d\gamma_t. \\ S_{0,0} &= 1 \end{aligned}$$

□

The solution to any linear equation is easily expressed in terms of the Signature

$$\begin{aligned} dy_t &= Ay_t d\gamma_t \\ y_{J_+} &= \left(\sum_0^{\infty} A^n S_J^n \right) y_{J_-} \\ \pi_J &= \sum_0^{\infty} A^n S_J^n \end{aligned} \tag{4.1}$$

and we will see in the next sections that this series converges very well and even the first few terms in S are effective in describing the response y_T leading to the view that $\gamma|_J \to S(\gamma|_J)$ is a transform with some value. The use of S to describe solutions to linear controlled differential equations goes back at least to Chen, and Feynman. The *magic* is that one can estimate the errors in convergence of the series (4.1) without detailed understanding of γ or A.

5. Remarkable estimates (for $p > 1$)

It seems strange, and even counter intuitive, that one should be able to identify and abstract a finite sequence of features or coefficients describing γ adequately so that its effect on a broad range of different systems could be accurately predicted without detailed knowledge of the system A or the path γ - beyond those few coefficients. But that is the truth of it, there are easy uniform estimates capturing the convergence of the series (4.1) based entirely on the length (or more generally p-rough path variation) of the control and the norm of A as a map from E to the linear vector fields on F.

Lemma 5.1. *If γ is a path of finite variation on J with length $|\gamma_J| < \infty$, then*

$$\begin{aligned} S_J^n &:= \int\cdots\int_{u_1\le\ldots\le u_n\in J^n} d\gamma_{u_1}\otimes\ldots\otimes d\gamma_{u_n} \\ &\le \frac{|\gamma_J|^n}{n!} \end{aligned}$$

giving uniform error control

$$\left\| y_{J_+} - \sum_0^{N-1} A^n \int\cdots\int_{J_-\le u_1\le\ldots\le u_n\le J_+} d\gamma_{u_1}\otimes\ldots\otimes d\gamma_{u_n} y_0 \right\| \le \left(\sum_{n=N}^{\infty} \frac{\|A\|^n |\gamma_J|^n}{n!} \right) \|y_0\| .$$

Proof. Because the Signature of the path always solves the characteristic differential equation it follows that one can reparameterise the path γ without changing the Signature of γ. Reparameterise γ so that it is defined on an interval J of length $|\gamma|$ and runs at unit speed. Now there are $n!$ disjoint simplexes inside a cube obtained by different permuted rankings of the coordinates and thus

$$\begin{aligned} \|S_J^n\| &:= \left\| \int\cdots\int_{u_1\le\ldots\le u_n\in J^n} d\gamma_{u_1}\otimes\ldots\otimes d\gamma_{u_n} \right\| \\ &= \left\| \int\cdots\int_{u_1\le\ldots\le u_n\in J^n} \dot\gamma_{u_1}\otimes\ldots\otimes\dot\gamma_{u_n} du_1\ldots du_n \right\| \\ &= \int\cdots\int_{u_1\le\ldots\le u_n\in J^n} \|\dot\gamma_{u_1}\otimes\ldots\otimes\dot\gamma_{u_n}\| \, du_1\ldots du_n \\ &= \int\cdots\int_{u_1\le\ldots\le u_n\in J^n} du_1\ldots du_n \\ &= \frac{|\gamma_J|^n}{n!}. \end{aligned}$$

from which the second estimate is clear. □

The Poisson approximation of a normal distribution one learns at high school ensures that the estimates on the right become very sharply estimated in terms of $\lambda\to\infty$ and pretty effective as soon as $N \ge \|A\|\,|\gamma_J| + \lambda\sqrt{\|A\|\,|\gamma_J|}$.

Remark 5.2. The uniform convergence of the series

$$\sum_{n=0}^{N-1} A^n \int\cdots\int_{J_-\le u_1\le\ldots\le u_n\le J_+} d\gamma_{u_1}\otimes\ldots\otimes d\gamma_{u_n} y_0$$

and the obvious continuity of the terms of the series in the inputs (A,γ,y_0) guarantees that the response y_T is jointly continuous (uniform limits of continuous functions are continuous) in (A,γ,y_0) where γ is given the topology of 1-variation (or any of the rough path metrics). It is already the case that

$$\gamma \to \int\cdots\int_{J_-\le u_1\le u_2\le J_+} d\gamma_{u_1}\otimes d\gamma_{u_2}$$

fails the closed graph property in the uniform metric.

6. The Log signature

It is easy to see that the Signature of a path segment actually takes its values in a very special curved subspace of the tensor algebra. Indeed, Chen noted that the map S is a homomorphism of path segments with concatenation into the algebra, and reversing the path segment produces the inverse tensor. As a result one sees that the range of the map is closed under multiplication and has inverses so it is a group (inside the grouplike elements) in the tensor series. It is helpful to think of the range of this Signature map as a curved space in the tensor series. As a result there is a lot of valuable structure. One important map is the logarithm; it is one to one on the group and provides a flat parameterisation of the group in terms of elements of the free Lie series.

Definition 6.1. If $\gamma_t \in E$ is a path segment and S is its Signature then

$$\begin{aligned} S &= 1 + S^1 + S^2 + \ldots \ \forall i,\ S^i \in E^{\otimes i} \\ \log(1+x) &= x - x^2/2 + \ldots \\ \log S &= \left(S^1 + S^2 + \ldots\right) - \left(S^1 + S^2 + \ldots\right)^2/2 + \ldots \end{aligned}$$

The series $\log S = \left(S^1 + S^2 + \ldots\right) - \left(S^1 + S^2 + \ldots\right)^2/2 + \ldots$ which is well defined, is referred to as the log Signature of γ.

Because the space of tensor series $T((E)) := \bigoplus_0^\infty E^{\otimes n}$ is a unital associative algebra under $\otimes, +$ it is also a Lie algebra, and with $[A,B] := A\otimes B - B\otimes A$.

Definition 6.2. There are several canonical Lie algebras associated to $T((E))$; we use the notation $\mathcal{L}(E)$ for the algebra generated by E (the space of Lie polynomials), $\mathcal{L}^{(n)}(E)$ the projection of this into $T^{(n)}(E) = T((E))/\bigoplus_{n+1}^\infty E^{\otimes m}$ (the Lie algebra of the free nilpotent group G^n of n steps) and $\mathcal{L}((E))$ the projective limit of the $\mathcal{L}^{(n)}(E)$ (the Lie Series).

Because we are working in characteristic zero, we may take the exponential, and this recovers the Signature, so no information is lost. A key observation of Chen [6]was that if

γ is a path segment then $\log S(\gamma) \in \mathcal{L}((E))$. The map from paths [8, 23]to $\mathcal{L}^{(n)}(E)$ via the projection $\pi_n : T((E)) \to T^{(n)}(E)$ is onto. Up to equivalence under a generalised notion of reparameterisation of paths known as treelike equivalence, the map from paths γ of finite length in E to their Signatures $S(\gamma) \in T((E))$ or log-Signatures $\log S \in \mathcal{L}((E))$ is injective [15]. Treelike equivalence is an equivalence relation on paths of finite variation, each class has a unique shortest element, and these tree reduced paths form a group. However the range of the log-Signature map in $\mathcal{L}((E))$, although well behaved under integer multiplication is not closed under integer division [21] and so the Lie algebra of the group of tree reduced paths is well defined but not a linear space; it is altogether a more subtle object.

Implicit in the definition of a controlled differential equation

$$dy_t = f(y_t)\, d\gamma_t,\ y_0 = a$$

is the map f. This object takes an element $e \in E$ and an element $y \in F$ and produces a second vector in F, representing the infinitesimal change to the state y of the system that will occur if γ is changed infinitesimally in the direction e. This author is clear that the best way to think about f is as a linear map from the space E into the vector fields on F. In this way one can see that the integral of f along γ in its simplest form is a path in the Lie algebra and that in solving the differential equation we are developing that path into the group. Now, at least formally, the vector fields are a Lie algebra (for the diffemorphisms of F) and subject to the smoothness assumptions we can take Lie brackets to get new vector fields. Because $\mathcal{L}((E))$ is the free Lie algebra over E (Chapter II, [2]) any linear map f of E into a Lie algebra $\mathfrak{g}$ induces a unique Lie map extension f_* to a Lie map from $\mathcal{L}((E))$ to $\mathfrak{g}$. This map can be readily implemented and is well defined because of the abstract theory

$$\begin{array}{rcl} e & \to & f(e) \quad \text{a vector field} \\ e_1e_2 - e_2e_1 & \to & f(e_1) f(e_2) - f(e_2) f(e_1) \quad \text{a vector field} \\ \tilde{f} & : & \mathcal{L}^{(n)}(E) \to \text{vector fields.} \end{array}$$

although in practice one does not take the map to the full projective limit.

7. The ODE method

The linkage between truncations of the log-Signature in $\mathcal{L}((E))$ and vector fields on Y is a practical one for modelling and understanding controlled differential equations. It goes well beyond theory and underpins some of the most effective and stable numerical approaches (and control mechanisms) for translating the information in the control γ into information about the response.

If $dy_t = f(y_t)\, d\gamma_t$, and $y_{J_-} = a$ then how can we use the first few terms of the (log-) Signature of γ to provide a good approximation to y_{J_+}? We could use picard iteration, or better an euler method based on a Taylor series in terms of the Signatures. Picard iteration for $\exp z$ already illustrates one issue. Picard interation yields a power series as approximation - fine if $z = 100$,but awful if $x = -100$. However, there is a more subtle problem to do with stability that almost all methods based on Taylor series have - stability - they can easily produce approximations that are not feasible. These are aggravated in the controlled case because of the time varying nature of the systems. It can easily happen that the solutions to the vector fields are hamiltonian etc. The ODE method uses the first few terms of the

Signature to construct a time invariant ODE (vector field) that if one solves it for unit time, it provides an approximation to the desired solution. It pushes the numerics back onto state of the art ODE solvers. Providing the ODE solver is accurate and stable then the approximation to y will also be. One can use symplectic solvers etc. At the level of rough paths, the approximation is obtained by replacing the path γ with a new rough path $\hat{\gamma}$ (a geodesic in the nilpotent group G^n) with the same first few terms in the Signature; this guarantees the feasibility of the approximations. Today, rough path theory can be used to estimate the difference between the solution and the approximation in terms of the distance between γ and $\hat{\gamma}$ even in infinite dimensions.[5][3]

Remark 7.1. A practical numerical scheme can be built as follows.

1. Describe γ over a short interval J in terms of first few terms of $\log S\left(\gamma_{[J_-,J_+]}\right)$ expressed as a linear combination of terms of a fixed hall basis:

$$\begin{aligned} \log S_J &= l^1 + l^2 + \ldots \in \mathcal{L}\left((E)\right) \\ l^{(n)} &= \pi_n\left(\log S_J\right) = l^1 + \ldots + l^n \in \mathcal{L}^{(n)}\left(E\right) \\ l^1 &= \sum_i \lambda_i e_i \\ l^2 &= \sum_{i<j} \lambda_{ij}\left[e_i, e_j\right], \\ &\cdots \end{aligned}$$

   and use this information to produce a path dependent vector field $V = \tilde{f}\left(l^{(n)}\right)$.
2. Use an appropriate ODE solver to solve the ODE $\dot{x}_t = V\left(x_t\right)$, where $x_0 = y_{J_-}$. A stable high order approximation to y_{J_+} is given by x_{J_+}.
3. Repeat over small enough time steps for the high order approximations to be effective.
4. The method is high order, stable, and corresponding to replacing γ with a piecewise geodesic path on successively finer scales.

8. Going to rough paths

As this is a survey, we have deliberately let the words rough path enter the text before they are introduced more formally. Rough path theory answers the following question. Suppose that γ is a smooth path but still on normal scales, a highly rough and oscillatory path. Suppose that we have some smooth system f . Give a simple metric on paths γ and a continuity estimate that ensures that if two paths that are close in this metric then their responses are quantifiably close as well. The estimate should only depend on f through its smoothness. There is such a theory [20], and a family of rough path metrics which make the function $\gamma \to y$ uniformly continuous. The completion of the smooth paths γ under these metrics are the rough paths we speak about. The theory extends to an infinite dimensional one and the estimates are uniform in a way that does not depend on dimension.

There are many sources for this information on rough paths for different kinds of audience and we do not repeat that material. We have mentioned that two smooth paths have quantifiable close responses to a smooth f over a fixed time interval if the first terms in the

Signature agree over this time interval. We can build this into a metric:

$$d_p\left(\gamma|_J, \hat{\gamma}|_J\right) = \sup_{J_- \le u_1 \le \ldots \le u_n \le J_+} \sum_i \max_{m \le \lfloor p \rfloor} \left\| S^m\left(\gamma|_{[u_i, u_{i+1}]}\right) - S^m\left(\hat{\gamma}|_{[u_i, u_{i+1}]}\right) \right\|^{p/m}$$

and providing the system is $Lip\left(p+\varepsilon\right)$ the response will behave uniformly with the control. The completion of the piecewise smooth paths under d_p are p-variation paths. They do not have smoothness but they do have a "top down" description and can be viewed as living in a $\lfloor p \rfloor$-step nilpotent group over E.

It is worth distinguishing the Kolmogorov and the rough path view on paths. In the former, one considers fixed times t_i, open sets O_i, and considers the probability that for all i, $x_{t_i} \in O_i$. In other words the emphasis is on where the path is at given times. This gated description will never capture the rough path; parameterisation is irrelevant but increments over small intervals $[u_i, u_{i+1}]$, are critical. More accurately one describes a path through an examination of the effect of it's path segment into a simple nonlinear system (the lift onto a nilpotent group). Knowing this information in an analytically adequate way is all one needs to know to predict the effect of the path on a general system.

The whole rough path theory is very substantial and we cannot survey it adequately here. The range is wide, and is related to any situation where one has a family of non-commuting operators and one wants to do analysis on apparently divergent products and for example it is interesting to understand the paths one gets as partial integrals of complex Fourier transform as the nonlinear Fourier transform is a differential equation driven by this path. Some results have been obtained in this direction [22] while the generalisations to spatial contexts are so huge that they are spoken about elsewhere at this congress. Many books are now written on the subject [11].and new lecture notes by Friz are to appear soon with recent developments. So in what is left of this paper we will focus on one topic the Signature of a path and the expected Signature of the path with a view to partially explaining how it is really an extension of Taylor's theorem to various infinite dimensional groups, and how we can get practical traction from this perspective. One key point we will not mention is that using Taylor's theorem twice works! This is actually a key point that the whole rough path story depends on and which validates its use. One needs to read the proofs to understand this adequately and, except for this sentence, suppress it completely here.

9. Coordinate iterated integrals

In this short paper we have to have a focus, and as a result we cannot explore the analysis and algebra needed to fully describe rough paths or to discuss the spatial generalisations directly even though they are having great impact[14][13]. Nonetheless much of what we say can be though of as useful foundations for this work. We are going to focus on the Signature as a tool for understanding paths and as a new tool to help with machine learning.

The essential remark may seem a bit daunting to an analyst, but will be standard to others. *The dual of the enveloping algebra of a group(like) object has a natural abelian product structure and linearises polynomial functions on a group.* This fact allows one to use linear techniques on the linear spaces to approximate generic smooth (and nonlinear) functions on the group. Here the group is the "group" of paths.

Monomials are special functions on $\mathbb{R}^n$, and polynomials are linear combinations of these monomials. Because monomials span an algebra, the polynomials are able to approx-

imate any continuous function on a compact set. Coordinate iterated integrals are linear functionals on the tensor algebra and at the same time they are the monomials or the features on path space.

Definition 9.1. Let $\boldsymbol{e} = e_1 \otimes \ldots \otimes e_n \in (E^*)^{\otimes n} \subset T(E^*)$, and $\phi_{\boldsymbol{e}}(\gamma) := \langle \boldsymbol{e}, S(\gamma) \rangle$ then we call $\phi_{\boldsymbol{e}}(\gamma)$ a coordinate iterated integral.

Remark 9.2. Note that $S(\gamma) \in T((E)) = \bigoplus_0^\infty E^{\otimes n}$ and

$$\begin{aligned} \phi_{\boldsymbol{e}}(\gamma) &= \langle \boldsymbol{e}, S(\gamma) \rangle \\ &= \int_{u_1 \leq \ldots \leq u_n \in J^n} \cdots \int \langle e_1, d\gamma_{u_1} \rangle \ldots \langle e_n, d\gamma_{u_n} \rangle \end{aligned}$$

justifying the name. $\phi_{\boldsymbol{e}}$ is a real valued function on Signatures of paths.

Lemma 9.3. *The shuffle product* $\amalg$ *on* $T(E^*)$ *makes* $T(E^*)$ *a commutative algebra and corresponds to point-wise product of coordinate integrals*

$$\phi_{\boldsymbol{e}}(\gamma)\, \phi_{\boldsymbol{f}}(\gamma) = \phi_{\boldsymbol{e} \amalg \boldsymbol{f}}(\gamma)$$

This last identity, which goes back to Ree, is important because it says that if we consider two linear functions on $T((E))$ and multiply them together then their product - which is quadratic actually agrees with a linear functional on the group like elements. The shuffle product identifies the linear functional that does the job.

Lemma 9.4. *Coordinate iterated integrals, as features of paths, span an algebra that separates Signatures and contains the constants.*

This lemma is as important for understanding smooth functions on path spaces as monomials are for understanding smooth functions on $\mathbb{R}^n$.There are only finitely many of each degree if E is finite dimensional (although the dimension of the spaces grow exponentially) [20]. We will see later that this property is important for machine learning and nonlinear regression applications but first we want to explain how the same remark allows one to understand measures on paths and formulate the notion of Fourier and Laplace transform.

10. Expected signature

The study of the expected Signature was initiated by Fawcett in his thesis [9]. He proved

Proposition 10.1. *Let* μ *be a compactly supported probability measure on paths* γ *with Signatures in a compact set* K*. Then* $\hat{S} = \mathbb{E}_\mu(S(\gamma))$ *uniquely determines the law of* $S(\gamma)$*.*

Proof. Consider $\mathbb{E}_\mu(\phi_{\boldsymbol{e}}(\gamma))$.

$$\begin{aligned} \mathbb{E}_\mu(\phi_{\boldsymbol{e}}(\gamma)) &= \mathbb{E}_\mu(\langle \boldsymbol{e}, S(\gamma) \rangle) \\ &= \langle \boldsymbol{e}, \mathbb{E}_\mu(S(\gamma)) \rangle \\ &= \left\langle \boldsymbol{e}, \hat{S} \right\rangle \end{aligned}$$

Since the $\boldsymbol{e}$ with the shuffle product form an algebra and separate points of K the Stone-Weierstrass Theorem implies they form a dense subspace in $C(K)$ and so determine the law of the Signature of γ. □

Given this lemma it immediately becomes interesting to ask how does one compute $\mathbb{E}_\mu(S)$. Also, $\mathbb{E}_\mu(S)$ is like a Laplace transform and will fail to exist for reasons of tail behaviour of the random variables. Is there a characteristic function? Can we identify the general case where the expected Signature determines the law in the non-compact case. All of these are fascinating and important questions. Partial answers and strong applications are emerging. One of the earliest was the realisation that one could approximate effectively to a complex measure such as Wiener measure by a measure on finitely many paths that has the same expected Signature on $T^{(n)}(E)$[17, 19].

11. Computing expected signatures

Computing Laplace and Fourier transforms can often be a challenging problem for undergraduates. In this case suppose that X a Brownian motion with Lévy area on a bounded C^1 domain $\Omega \subset \mathbb{R}^d$,stopped on first exit. The following result explains how one may construct the expected Signature as a recurrence relation in PDEs[18].

Theorem 11.1. *Let*

$$\begin{aligned} F(z) &: = \mathbb{E}_z\left(S\left(X|_{[0,T_\Omega]}\right)\right) \\ F &\in S\left(\left(\mathbb{R}^d\right)\right) \\ F &= (f_0, f_1, \ldots,) \end{aligned}$$

Then F satisfies and is determined by a PDE finite difference operator

$$\begin{aligned} \Delta f_{n+2} &= -\sum_{i=1}^{d} e_i \otimes e_i \otimes f_n - 2\sum_{i=1}^{d} e_i \otimes \frac{\partial}{\partial z_i} f_{n+1} \\ f_0 &\equiv 1,\ f_1 \equiv 0, \text{ and } f_j|_{\partial\Omega} \equiv 0,\ j > 0 \end{aligned}$$

Combining this result with Sobolev and regularity estimates from PDE theory allow one to extract much nontrivial information about the underlying measure although it is still open whether in this case the expected Signature determines the measure. This question is difficult even for Brownian motion on $\min(T_\tau, t)$ although (unpublished) it looks as if the question can be resolved.

Other interesting questions about expected Signatures can be found for example in [1].

12. Characteristic functions of signatures

It is possible to build a characteristic function out of the expected Signature by looking at the linear differential equations corresponding to development of the paths into finite dimensional unitary groups. These linear images of the Signature are always bounded and so expectations always make sense.

Consider $SU(d) \subset M(d)$ and realise $su(d)$ as the space of traceless Hermitian matrices and consider

$$\psi : E \to su(d)$$

$$d\Psi_t \quad = \quad \psi\left(\Psi_t\right) d\gamma_t.$$

Essential features of the co-ordinate interated integrals included that they were linear functions on the tensor algebra, that they were real valued functions that separated signatures, and that they spanned an algebra.

It is core to rough path theory that any representation of paths via a linear controlled equation can also be regarded as a linear function and that products can also be represented as sums. If one can show that products associated to the finite dimensional unitary groups can be expressed as sums of finite linear combinations of finite dimensional unitary representations, and add an appropriate topology on grouplike elements, one can repeat the ideas outlined above but now with expectations that always exist and obtain the analogue of characteristic function.

Theorem 12.1. *Ψ_t is a linear functional on the tensor algebra restricted to the Signatures $S\left(\gamma|_{[0,t]}\right)$ and is given by a convergent series. It is bounded and so its expectation as γ varies randomly always makes sense. The function $\psi \to \mathbb{E}\left(\Psi_{J_+}\left(S\right)\right)$ is an extended characteristic function.*

Proposition 12.2. *$\psi \to \Psi\left(S\right)$ (polynomial identities of Gambruni and Valentini) span an algebra and separate Signatures as ψ and d vary.*

Corollary 12.3. *The laws of measures on Signatures are completely determined by $\psi \to \mathbb{E}\left(\Psi\left(S\right)\right)$*

Proof. Introduce a polish topology on the grouplike elements. □

These results can be found in [7], the paper also gives a sufficient condition for the expected Signature to determine the law of the underlying measure on Signatures.

13. Moments are complicated

The question of determining the Signature from its moments seems quite hard at the moment.

Example 13.1. Observe that if X is $N\left(0,1\right)$ then although X^3 is not determined by its moments, if $Y = X^3$ then $\left(X,Y\right)$ is. The moment information implies $\mathbb{E}\left(\left(Y-X^3\right)^2\right) = 0$.

We repeat our previous question. Does the expected Signature determine the law of the Signature for say stopped Brownian motion. The problem seems to capture the challenge.

Lemma 13.2 ([7]). *If the radius of convergence of $\sum z^n \mathbb{E}\left\|S^n\right\|$ is infinite then the expected Signature determines the law.*

Lemma 13.3 ([18]). *If X a Brownian motion with Lévy area on a bounded C^1 domain $\Omega \subset \mathbb{R}^d$ then $\sum z^n \mathbb{E}\left\|S^n\right\|$ has at the least a strictly positive lower bound on the radius of curvature.*

The gap in understanding between the previous two results is, for the author, a fascinating and surprising one that should be closed!

14. Regression onto a feature set

Learning how to regress or learn a function from examples is a basic problem in many different contexts. In what remains of this paper, we will outline recent work that explains how the Signature engages very naturally with this problem and why it is this engagement that makes it valuable in rough path theory too.

We should emphasise that the discussion and examples we give here is at a very primitive level of fitting curves. We are not trying to do statistics, or model and make inference about uncertainty. Rather we are trying to solve the most basic problems about extracting relationships from data that would exist even if one had perfect knowledge. We will demonstrate that this approach can be easy to implement and effective in reducing dimension and doing effective regression. We would expect Baysian statistics to be an added layer added to the process where uncertanty exists in the data that can be modelled reasonably.

A core idea in many successful attempts to learn functions from a collection of known (point, value) pairs revolves around the identification of basic functions or features that are readily evaluated at each point and then try to express the observed function a *linear* combination of these basic functions. For example one might evaluate a smooth function ρ at a generic collection $\{x_i \in [0,1]\}$ of points producing pairs $\{(y_i = \rho(x_i), x_i)\}$ Now consider as feature functions $\{\phi_n : x \to x^n, n = 0, \dots N\}$. These are certainly easy to compute for each x_i. We try to express

$$\rho \simeq \sum_{n=0}^{N} \lambda_n \phi_n$$

and we see that if we can do this (that is to say ρ is well approximated by a polynomial) then the λ_n are given by the linear equation

$$y_j = \sum_{n=0}^{N} \lambda_n \phi_n(x_j).$$

In general one should expect, and it is even desirable, that the equations are significantly degenerate. The purpose of learning is presumably to be able to use the function $\sum_{n=0}^{N} \lambda_n \phi_n$ to predict ρ on new and unseen values of x and to at least be able to replicate the observed values of y.

There are powerful numerical techniques for identifying robust solutions to these equations. Most are based around least squares and singular value decomposition, along with L^1 constraints and Lasso.

However, this approach fundamentally depends on the assumption that the ϕ_n span the class of functions that are interesting. It works well for monomials because they span an algebra and so every $C^n(K)$ function can be approximated in $C^n(K)$ by a multivariate real polynomial. It relies on a priori knowledge of smoothness or Lasso style techniques to address over-fitting.

I hope the reader can now see the significance of the coordinate iterated integrals. If we are interested in functions (such as controlled differential equations) that are effects of paths or streams, then we know from the general theory of rough paths that the functions are indeed well approximated locally by linear combinations of coordinate iterated integrals. Coordinate iterated integrals are a natural feature set for capturing the aspects of the data that predicting the effects of the path on a controlled system.

The shuffle product ensures that linear combinations of coordinate iterated integrals are an algebra which ensures they span adequately rich classes of functions. We can use the classical techniques of non-linear interpolation with these new feature functions to learn and model the behaviour of systems.

In many ways the machine learning perspective explains the whole theory of rough paths. If I want to model the effect of a path segment, I can do a good job by studying a few set features of my path locally. On smaller scales the approximations improve since the functionals the path interacts with become smoother. If the approximation error is small compared with the volume, and consistent on different scales, then knowing these features, and only these features, on all scales describes the path or function adequately enough to allow a limit and integration of the path or function against a Lipchitz function.

15. The obvious feature set for streams

The feature set that is the coordinate iterated integrals is able (with uniform error - even in infinite dimension) via linear combinations whose coefficients are derivatives of f, to approximate solutions to controlled differential equations [3]. In other words, any stream of finite length is characterised up to reparameterisation by its log Signature (see [15]) and the Poincare-Birkhoff-Witt theorem confirms that the coordinate iterated integrals are one way to parameterise the polynomials on this space. Many important nonlinear functions on paths are well approximated by these polynomials.

We have a well defined methodology for linearisation of smooth functions on unparameterised streams as linear functionals of the Signature. As we will explain in the remaining sections, this has potential for practical application even if it comes from the local embedding of a group into its enveloping algebra and identifying the dual with the real polynomials and analytic functions on the group.

16. Machine learning, an amateur's first attempt

Applications do not usually have a simple fix but require several methods in parallel to achieve significance. The best results to date for the use of Signatures have involved the recognition of Chinese characters [24] where Ben Graham put together a set of features based loosely on Signatures and state of the art deep learning techniques to win a worldwide competition organised by the Chinese Academy of Sciences.

We will adopt a different perspective and simply explain a very transparent and naive approach, based on Signatures, can achieve with real data. The work appeared in [12]. The project and the data depended on collaboration with commercial partners acknowledged in the paper and is borrowed from the paper.

16.1. classification of time-buckets from standardised data. We considered a simple classification learning problem. We considered a moderate data set of 30 minutes intervals of normalised one minute financial market data, which we will call buckets. The buckets are distinguished by the time of day that the trading is recorded. The buckets are divided into two sets - a learning and a backtesting set. The challenge is simple: learn to distinguish

the time of day by looking at the normalised data (if indeed one can - the normalisation is intended to remove the obvious). It is a simple classification problem that can be regarded as learning a function with only two values

$$\begin{array}{ccc} f\,(\text{time series}) & \rightarrow & \text{time slot} \\ f\,(\text{time series}) = 1 & & \text{time slot=10.30-11.00} \\ f\,(\text{time series}) = 0 & & \text{time slot=14.00-14.30} \end{array} .$$

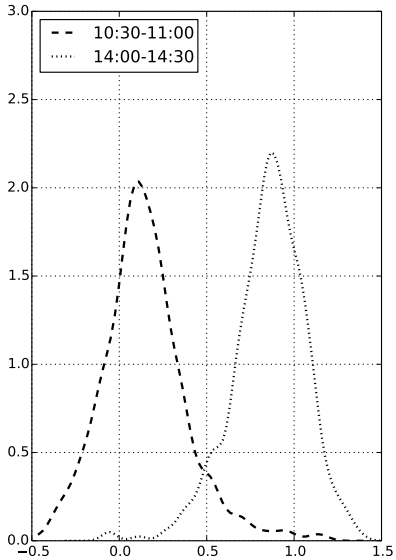
Our methodology has been spelt out. Use the low degree coordinates of the Signature of the normalised financial market data γ as features $\phi_i(\gamma)$, use least squares on the learning set to approximately reproduce f

$$f(\gamma) \approx \sum_i \lambda_i \phi_i(\gamma)$$

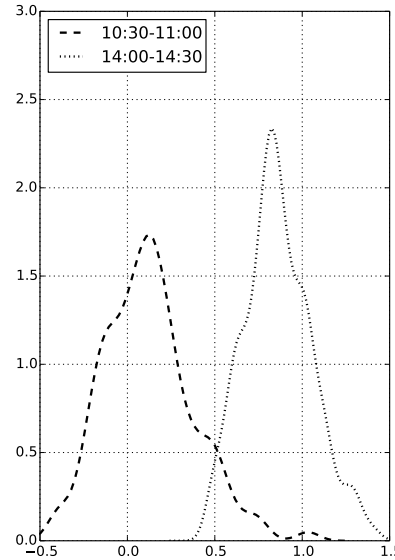
and then test it on the backtesting set. To summarise the methodology:

1. We used futures data normalised to remove volume and volatility information.
2. We used linear regression based pair-wise separation to find the best fit linear function to the learning pairs that assign 0 to one case and 1 to the other. (There are other well known methods that might be better.)
   (a) We used robust and automated repeated sampling methods of LASSO type (least absolute shrinkage and selection operator) based on constrained L^1 optimisation to achieve shrinkage of the linear functional onto an expression involving only a few terms of the Signatures.
3. and we used simple statistical indicators to indicate the discrimination that the learnt function provided on the learning data and then on the backtesting data. The tests were:
   (a) Kolmogorov-Smirnov distance of distributions of score values
   (b) receiver operating characteristic (ROC) curve, area under ROC curve
   (c) ratio of correct classification.

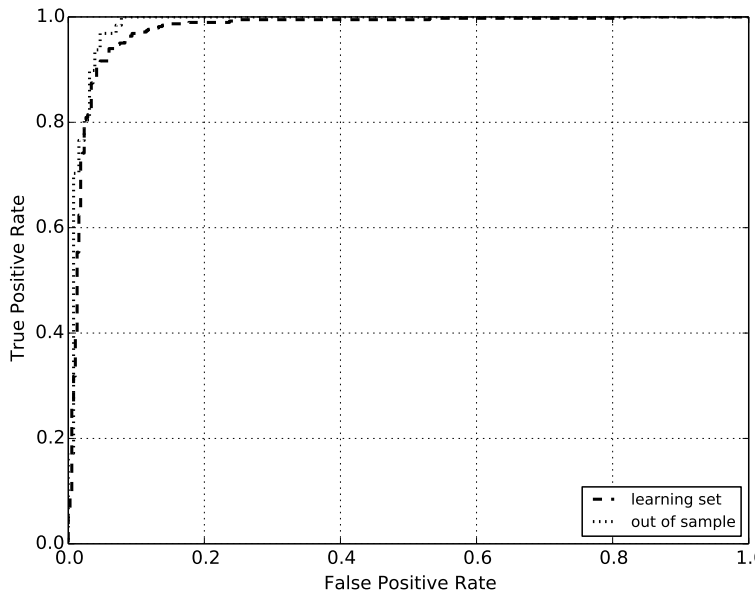
We did consider the full range of half hour time intervals. The other time intervals were not readily distinguishable from each other but were easily distinguishable from both of these two time intervals using the methodology mapped out here. It seems likely that the differences identified here were due to distinctive features of the market associated with the opening and closing of the open outcry market.



(a) **Learning set**: Estimated densities of the regressed values, K-S distance: 0.8, correct classification: 90%



(b) **Out of sample**: Estimated densities of the regressed values, K-S distance: 0.84, correct classification: 89%



(c) **ROC curve.** Area under ROC – learning set: 0.976, out of sample: 0.986

Figure 16.1. 14:00-14:30 EST versus 10:30-11:00 EST

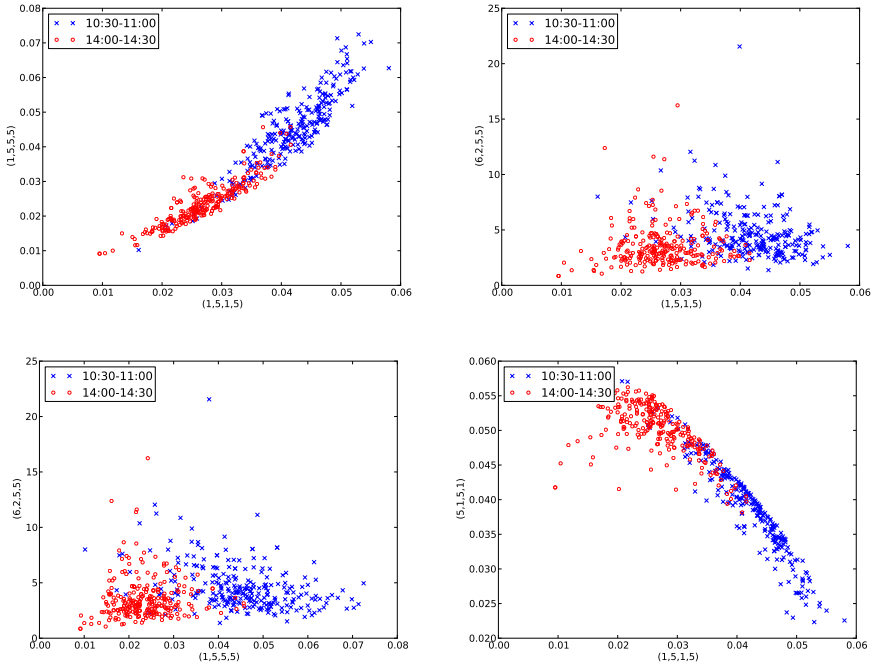


Figure 16.2. Visualisation: two dimensional projections of the 4th order signature onto coefficients selected as significant by Lasso shrinkage. The selected features allow clear visual separation of the time buckets.

17. Linear regression onto a law on paths

In the previous section we looked at using the linearising nature of the Signature as a pratical tool for learning functions. In this final section we want to remain in the world of data and applications but make a more theoretical remark. Classic nonlinear regression is usually stated with a statistical element. One common formulation of linear regression has that a stationary sequence of random data pairs that are modeled by

$$y_i = f(x_i) + \varepsilon_i$$

where ε_i is random and has conditional mean zero. The goal is to determine the linear functional f with measurable confidence.

There are many situations where it is the case that one has a random but stationary sequence (γ, τ) of stream pairs, and one would like to learn, approximately, the law of τ conditional on γ. Suppose that we reformulate this problem in terms of Signatures and expected Signatures (or better: charateristic functions) recalling that expected Signatures etc. characterise laws.

Problem 17.1. Given a random but stationary sequence (γ, τ) of stream pairs find the function $\Phi : S(\gamma) \to \mathbb{E}(S(\tau) | S(\gamma))$.

Then putting $Y_i = S(\tau_i)$ and $X_i = S(\gamma_i)$ we see that

$$Y_i = \Phi(X_i) + \varepsilon_i$$

where ε_i is random and has mean zero. If the measure is reasonably localised and smooth then we can well approximate Φ by a polynomial; and using th elinearising nature of the tensor algebra to a linear function ϕ of the Signature. In other words the apparently difficult problem of understanding conditional laws of paths becomes (at least locally) a problem of linear regression

$$Y_i = \Phi(X_i) + \varepsilon_i$$

whch is infinite dimensional but which has well defined low dimensional approximations [16].

Acknowledgement. The author acknowledges the support provided by ERC advanced grant ESig (agreement no. 291244), and particularly the contributions of his colleagues and his students without whom none of this would have happened and in addition to Kelly Wyatt, Justin Sharp, Horatio Boedihardjo, Hao Ni and Danyu Yang for helping the author finalise this mss. The data analysis is is reproduced from the cited paper with Gyurko et al., Gyurko did the analysis. The raw data for that paper is available on Reuters.

References

[1] Horatio Boedihardjo, Hao Ni, and Zhongmin Qian, *Uniqueness of signature for simple curves*, ArXiv preprint arXiv:1304.0755 (2013), 1–21.

[2] Nicolas Bourbaki, *Lie groups and Lie algebras. Chapters 1–3*, Elements of Mathematics (Berlin), Springer-Verlag, Berlin, 1989, Translated from the French, Reprint of the 1975 edition. MR 979493 (89k:17001)

[3] Youness Boutaib, Lajos Gergely Gyurkó, Terry Lyons, and Danyu Yang, *Dimension-free euler estimates of rough differential equations*, arXiv:1307.4708 to appear in Rev. Roumaine Math. Pures Appl. (2014), 1–20.

[4] Thomas Cass and Terry Lyons, *Evolving communities with individual preferences*, 1303.4243 to appear in Proceedings of London Mathematical Society (2014), 1–21.

[5] Fabienne Castell and Jessica Gaines, *An efficient approximation method for stochastic differential equations by means of the exponential lie series*, Mathematics and computers in simulation **38** (1995), no. 1, 13–19.

[6] Kuo-Tsai Chen, *Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula*, Ann. of Math. (2) **65** (1957), 163–178. MR 0085251 (19,12a)

[7] Ilya Chevyrev, *Unitary representations of geometric rough paths*, arXiv preprint arXiv: 1307.3580 (2014).

[8] Wei-Liang Chow, *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann. **117** (1939), 98–105. MR 0001880 (1,313d)

[9] Thomas Fawcett, *Problems in stochastic analysis: Connections between rough paths and non-commutative harmonic analysis*, Ph.D. thesis, University of Oxford, 2002.

[10] Guy Flint, Ben Hambly, and Terry Lyons, *Convergence of sampled semimartingale rough paths and recovery of the itô integral*, arXiv preprint arXiv:1310.4054v5 (2013), 1–22.

[11] Peter K Friz and Nicolas B Victoir, *Multidimensional stochastic processes as rough paths: theory and applications*, vol. 120, Cambridge University Press, 2010.

[12] Lajos Gergely Gyurkó, Terry Lyons, Mark Kontkowski, and Jonathan Field, *Extracting information from the signature of a financial data stream*, arXiv preprint arXiv: 1307.7244 (2013).

[13] Martin Hairer, *A theory of regularity structures*, Invent. Math. (2014).

[14] Martin Hairer and Natesh S Pillai, *Regularity of laws and ergodicity of hypoelliptic sdes driven by rough paths*, The Annals of Probability **41** (2013), no. 4, 2544–2598.

[15] Ben Hambly and Terry Lyons, *Uniqueness for the signature of a path of bounded variation and the reduced path group*, Ann. of Math.(2) **171** (2010), no. 1, 109–167.

[16] Daniel Levin, Terry Lyons, and Hao Ni, *Learning from the past, predicting the statistics for the future, learning an evolving system*, arXiv preprint arXiv:1309.0260 (2013), 1–32.

[17] Christian Litterer and Terry Lyons, *Cubature on wiener space continued*, Stochastic Processes and Applications to Mathematical Finance (2011), 197–218.

[18] Terry Lyons and Hao Ni, *Expected signature of two dimensional Brownian Motion up to the first exit time of the domain*, arXiv:1101.5902v4 (2011), 1–27.

[19] Terry Lyons and Nicolas Victoir, *Cubature on wiener space*, Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **460** (2004), no. 2041, 169–198.

[20] Terry J Lyons, Michael Caruana, and Thierry Lévy, *Differential equations driven by rough paths*, Springer, 2007.

[21] Terry J. Lyons and Nadia Sidorova, *On the radius of convergence of the logarithmic signature*, Illinois J. Math. **50** (2006), no. 1-4, 763–790 (electronic). MR 2247845 (2007m:60165)

[22] Terry J Lyons and Danyu Yang, *The partial sum process of orthogonal expansions as geometric rough process with fourier series as an exampleâĂŤan improvement of menshov–rademacher theorem*, Journal of Functional Analysis **265** (2013), no. 12, 3067–3103.

[23] P. K. Rashevski, *About connecting two points of complete nonholonomic space by admissible curve*, Uch Zapiski ped. inst. Libknekhta **2** (1938), 83–94.

[24] Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, and Cheng-Lin Liu, *Icdar 2013 chinese handwriting recognition competition*, Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, IEEE, 2013, pp. 1464–1470.

Oxford-Man Institute of Quantitative Finance, University of Oxford, England, OX2 6ED
E-mail: terry.lyons@oxford-man.ox.ac.uk

Variational formulas for directed polymer and percolation models

Timo Seppäläinen

Abstract. Explicit formulas for subadditive limits of polymer and percolation models in probability and statistical mechanics have been difficult to find. We describe variational formulas for these limits and connections with other features of the models such as Busemann functions and Kardar-Parisi-Zhang (KPZ) fluctuation exponents.

Mathematics Subject Classification (2010). Primary 60K35; Secondary 60K37, 82B41.

Keywords. Busemann function, cocycle, convex duality, directed polymer, last-passage percolation, Kardar-Parisi-Zhang universality, random environment, variational formula.

1. Introduction

This paper reviews recently discovered variational formulas for the limiting free energy density and time constant of directed polymer and percolation models, together with connections to other features of the models such as Busemann functions and fluctuation exponents. The existence of these limits comes from subadditive ergodic theory. Consequently there is no obvious formula for the limit, in contrast with the additive ergodic theorem whose limit is an expectation. We restrict the discussion to the simplest path geometry in the plane. Some results generalize in various directions, such as higher dimension, more general ergodic environment, and more general admissible paths. The results outlined in this note come from papers [12, 13, 27, 28, 32].

Busemann functions, as limits of gradients of passage times, were introduced into first-passage percolation by C. Newman and coworkers in the 1990's [25]. Since then Busemann functions have emerged as a central object in the study of geodesics and invariant distributions of percolation models and related interacting particle systems. See [3, 7, 11, 17] for a selection of the literature. In this review we look at Busemann functions for the log-gamma polymer in the positive temperature setting, as opposed to the zero-temperature setting of percolation. The connection with the variational formulas comes from the fact that the Busemann functions are minimizers in a variational problem that characterizes the limiting free energy density of the log-gamma polymer.

Directed percolation and polymer models on the plane are expected to be members of the Kardar-Parisi-Zhang (KPZ) universality class. The name comes from the influential 1986 paper [21]. For a mathematical review, see [9]. The KPZ universality class is characterized by the following features:

Proceedings of the International Congress of Mathematicians, Seoul, 2014

(i) In a system of size $n \times n$, fluctuations of quantities such as the free energy and the passage time have order of magnitude $n^{1/3}$ while fluctuations of polymer paths have order of magnitude $n^{2/3}$.
(ii) In the $n \to \infty$ limit, properly scaled random quantities converge weakly to limit distribution that come from random matrix theory.

This should be contrasted with the diffusive behavior of random walk: a path of length n fluctuates on the scale $n^{1/2}$ and limit distributions are Gaussian. KPZ universality remains presently a conjecture. Results on both the fluctuation exponents and limit distributions exist for a handful of exactly solvable models [2, 5, 6, 9, 20, 26, 33, 34].

Fluctuation behavior is not the focus of this paper, but we will touch upon KPZ fluctuation exponents at the end. The exponents arise naturally in this context because they can be studied through controlling the fluctuations of Busemann functions. This can be achieved to satisfaction in exactly solvable models where the curvature of the limit shape is known.

Independently and simultaneously with the work described in Section 3, A. Krishnan derived an analogous variational formula for first-passage percolation where paths are not restricted to be directed [22].

Notation and some definitions. The ℓ^1 norm of a point $x = (x_1, x_2) \in \mathbb{R}^2$ is $|x|_1 = |x_1| + |x_2|$. The basis vectors of $\mathbb{R}^2$ are $e_1 = (1, 0)$ and $e_2 = (0, 1)$. The usual gamma function for $\rho > 0$ is $\Gamma(\rho) = \int_0^\infty x^{\rho-1}e^{-x}\,dx$. Random variable X has the Gamma(ρ) probability distribution on $\mathbb{R}_+$ if $P(X \le t) = \int_0^t \Gamma(\rho)^{-1}x^{\rho-1}e^{-x}\,dx$ for $t \ge 0$. A shorthand for this is $X \sim$ Gamma(ρ). $\Psi_0 = \Gamma'/\Gamma$ and $\Psi_1 = \Psi_0'$ are the digamma and trigamma functions.

2. Last-passage percolation and directed polymers on the planar lattice

We begin with a description of the probability space that contains the random weights used to construct the models. Let $\Omega = \mathbb{R}^{\mathbb{Z}^2}$ be the space of weight configurations $\omega = (\omega_x)_{x\in\mathbb{Z}^2}$. Ω is endowed with its Borel σ-algebra $\mathcal{S}$ and the group of translations $\{T_x\}_{x\in\mathbb{Z}^2}$ that act via $(T_x\omega)_y = \omega_{x+y}$ for $x, y \in \mathbb{Z}^2$. Let $\mathbb{P}$ be a Borel probability measure on $(\Omega, \mathcal{S})$ under which the weights ω_x are independent and identically distributed (i.i.d.) random variables. This means that, for any distinct vertices $x_1, \dots, x_n \in \mathbb{Z}^2$ and arbitrary Borels sets $A_1, \dots, A_n \subseteq \mathbb{R}$,

$$\mathbb{P}\{\omega_{x_1} \in A_1, \dots, \omega_{x_n} \in A_n\} = \prod_{i=1}^{n} \mathbb{P}\{\omega_0 \in A_i\}.$$

Such a measure $\mathbb{P}$ is invariant and ergodic under the translations. This means that $\mathbb{P}(T_xA) = \mathbb{P}(A)$ for every Borel set $A \subseteq \Omega$, and furthermore, an invariant Borel set A (one that satisfies $T_xA = A$ for all $x \in \mathbb{Z}^2$) has probability $\mathbb{P}(A) \in \{0, 1\}$. The probability space $(\Omega, \mathcal{S}, \mathbb{P})$ now contains a random weight ω_x assigned to each point $x \in \mathbb{Z}^2$. This provides the *random environment* or *random medium* ω.

Let $\Pi_{0,v}$ denote the set of directed lattice paths from the origin 0 to a fixed point v. A directed, or up-right, path is only allowed steps e_1 and e_2. The figure gives an example of a directed path from 0 to $v = (5, 4)$.

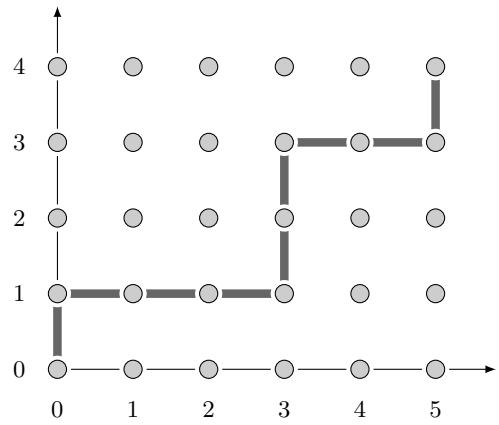


Figure 2.1. An up-right path from $0 = (0, 0)$ to $(5, 4)$ in $\mathbb{Z}^2_+$.

Fix a Borel function $V : \mathbb{R} \to \mathbb{R}$. The *directed polymer* model gives each directed path $x_\cdot = (x_0, \dots, x_{|v|_1})$ from 0 to v a probability proportional to the exponential of the weight of the path:

$$Q_{0,v}(x_\cdot) = \frac{1}{Z_{0,v}} 2^{-|v|_1} \exp\Big(\beta \sum_{k=0}^{|v|_1-1} V(\omega_{x_k})\Big) \qquad \text{for } x_\cdot \in \Pi_{0,v} \tag{2.1}$$

where

$$Z_{0,v} = \sum_{x_\cdot \in \Pi_{0,v}} 2^{-|v|_1} \exp\Big(\beta \sum_{k=0}^{|v|_1-1} V(\omega_{x_k})\Big) \tag{2.2}$$

is the *partition function* that normalizes $Q_{0,v}$ to a probability measure. The factor $2^{-|v|_1}$ is included only to make the sum into an expectation over equally likely random walk paths. The parameter $\beta > 0$ is the inverse temperature. The free energy $G^\beta_{0,v} = \beta^{-1} \log Z_{0,v}$ is a quantity of key interest. Note that $Q_{0,v}(x_\cdot)$, $Z_{0,v}$ and $G^\beta_{0,v}$ are all functions of the environment ω. For this reason $Q_{0,v}(x_\cdot)$ is called the *quenched* probability on paths and $Z_{0,v}$ the *quenched* partition function. We use the function $V(\omega_x)$ in the exponential, instead of simply the weight ω_x, to give us some added flexibility.

The model defined by (2.1)–(2.2) is called the *point-to-point* polymer because the path is pinned or fixed at both endpoints 0 and v. The analogous *point-to-line* model admits all paths of fixed length N that start at the origin. The set of paths is $\Pi_{0,(N)} = \cup_{|v|_1=N} \Pi_{0,v}$. In the point-to-line case it is natural to include a tilt or external field $h \in \mathbb{R}^2$ in the model. The quenched probability, partition function and free energy are

$$Q_{(N)}(x_\cdot) = \frac{1}{Z_{(N)}} 2^{-N} \exp\Big\{\beta \sum_{k=0}^{N-1} V(\omega_{x_k}) + \beta h \cdot x_N\Big\} \qquad \text{for } x_\cdot \in \Pi_{0,(N)}, \tag{2.3}$$

$$Z_{(N)} = \sum_{x_\cdot \in \Pi_{0,(N)}} 2^{-N} \exp\Big\{\beta \sum_{k=0}^{N-1} V(\omega_{x_k}) + \beta h \cdot x_N\Big\} \tag{2.4}$$

and

$$G^\beta_{(N)}(h) = \beta^{-1} \log Z_{(N)}. \tag{2.5}$$

The point-to-line case is distinguished by the subscript (N) in parentheses. While all the quantities depend on β, β is included explicitly only in the notation G^β because that is where we want to indicate explicitly the distinction between $\beta < \infty$ and $\beta = \infty$.

If we take $\beta \to \infty$ the probability measures $Q_{0,v}$ and $Q_{(N)}$ concentrate on those paths that maximize the exponent. This is the *zero-temperature polymer*, also known as the *last-passage percolation model* or the *corner growth model*. The key quantity is the maximal last-passage time of directed paths: in the point-to-point case

$$G^\infty_{0,v} = \max_{x_\cdot \in \Pi_{0,v}} \sum_{k=0}^{|v|_1 - 1} V(\omega_{x_k}) \tag{2.6}$$

and in the point-to-line case

$$G^\infty_{(N)}(h) = \max_{x_\cdot \in \Pi_{0,(N)}} \Big\{ \sum_{k=0}^{N-1} V(\omega_{x_k}) + h \cdot x_N \Big\}.$$

The directed percolation model admits a natural description as a model of a randomly growing cluster $\mathcal{A}_t$ on the plane. Suppose $V(\omega_x) \geq 0$ so that $V(\omega_x)$ can be interpreted as a waiting time. Then define the cluster at time $t \geq 0$ by

$$\mathcal{A}_t = \{x \in \mathbb{Z}^2_+ : G^\infty_{0,x} \leq t\}$$

The questions of interest are about the large-scale behavior of this model: namely, how do the random path and the quantities $G^\beta_{0,x}$ and $G^\beta_{(N)}(h)$ behave as the point x or the parameter N is taken to infinity?

Let us make a further assumption: namely, that the random weights satisfy a moment bound

$$\mathbb{E}(|V(\omega_0)|^p) < \infty \quad \text{for some } p > 2. \tag{2.7}$$

The starting points are the following laws of large numbers, for both $\beta < \infty$ and $\beta = \infty$: in the point-to-point case

$$g^\beta_{\text{pp}}(\xi) = \lim_{N\to\infty} N^{-1} G^\beta_{0,\lfloor N\xi \rfloor} \qquad \text{for } \xi \in \mathbb{R}^2_+ \tag{2.8}$$

and in the point-to-line case

$$g^\beta_{\text{pl}}(h) = \lim_{N\to\infty} N^{-1} G^\beta_{(N)}(h) \qquad \text{for } h \in \mathbb{R}^2. \tag{2.9}$$

The limits g^β_{pp} and g^β_{pl} are deterministic continuous functions of their arguments ξ and h. For the polymer model these are the limiting free energy densities and for the percolation

model the limiting time constants. In particular, g_{pp}^{∞} describes the limit shape of the growing cluster. As $t \to \infty$, the scaled random set $t^{-1}\mathcal{A}_t$ converges almost surely to the set $\{\xi \in \mathbb{R}^2_+ : g_{\text{pp}}^{\infty}(\xi) \le 1\}$. In first-passage percolation this *shape theorem* goes back to [10], and for last-passage percolation it is recorded in [23].

The function g_{pp}^{β} is homogeneous: $g_{\text{pp}}^{\beta}(c\xi) = cg_{\text{pp}}^{\beta}(\xi)$ for $c > 0$. Consequently it is sufficient and convenient to consider it only on the set $\mathcal{U} = \{\xi = (s, 1-s) : 0 \le s \le 1\}$ of asymptotic velocities of admissible paths. The relative interior of $\mathcal{U}$ is $\mathcal{U}^{\circ} = \{\xi = (s, 1-s) : 0 < s < 1\}$.

Once a law of large numbers is understood, the next questions asked in probability concern the chances of deviations. As mentioned in the Introduction, it is expected that, according to KPZ universality, deviations are of order $N^{1/3}$. This means that we would expect probabilities

$$\mathbb{P}\bigl\{G^{\beta}_{0,\lfloor N\xi\rfloor} - Ng^{\beta}_{\text{pp}}(\xi) \le N^{1/3}s\bigr\}$$

of fluctuations in the $N^{1/3}$ scale to converge to something nontrivial. Such results exist for exactly solvable models.

This section concludes with some historical remarks. The directed polymer model was introduced in 1985 by Huse and Henley [18]. First-passage percolation arose already in 1965 in the work of Hammersley and Welsh [15]. The origins of last-passage percolation are murkier. The corner growth model with exponential weights appeared in Rost's seminal paper [29] on a hydrodynamic limit of the totally asymmetric simple exclusion process, but without the last-passage formulation. The study of directed last-passage percolation picked up in the 1990's, with explicit shape results for exactly solvable cases in different geometries in the articles [1, 8, 19, 30, 31]. Early motivation for [1] came from [16].

3. Variational formulas for the limits

Properties of the limits $g_{\text{pp}}^{\beta}(\xi)$ and $g_{\text{pl}}^{\beta}(h)$ have remained difficult to analyze and, prior to the results explained below, no formulas for these limits have been found. Part of the reason is that these are instances of superadditive convergence. This can be contrasted with the classical law of large numbers or the ergodic theorem: if $\{X_k\}_{k\in\mathbb{N}}$ is a stationary, ergodic sequence of integrable random variables, then

$$\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} X_k = E(X_1) \qquad \text{almost surely} \tag{3.1}$$

and so there is no mystery about the limit value.

We proceed to describe variational formulas for the limits $g_{\text{pp}}^{\beta}(\xi)$ and $g_{\text{pl}}^{\beta}(h)$. Assumption (2.7) remains in force throughout the discussion. We need some preliminaries first.

Definition 3.1. A measurable function $B : \Omega \times \mathbb{Z}^2 \times \mathbb{Z}^2 \to \mathbb{R}$ is a *stationary* $L^1(\mathbb{P})$ *cocycle* if it satisfies the following three conditions.

(i) Integrability: for each $z \in \{e_1, e_2\}$, $\mathbb{E}|B(0,z)| < \infty$.

(ii) Stationarity: for $\mathbb{P}$-a.e. ω and all $x, y, z \in \mathbb{Z}^2$, $B(\omega, z+x, z+y) = B(T_z\omega, x, y)$.

(iii) Cocycle property: for $\mathbb{P}$-a.e. ω and for all $x, y, z \in \mathbb{Z}^2$, $B(\omega, x, y) + B(\omega, y, z) = B(\omega, x, z)$. The space of cocycles is denoted by $\mathcal{C}$.

A *centered cocycle* is a stationary L^1 cocycle $F(\omega, x, y)$ that satisfies $\mathbb{E}[F(x, y)] = 0$ for all $x, y \in \mathbb{Z}^2$. The space of centered cocycles is denoted by $\mathcal{C}_0$.

The space $\mathcal{C}_0$ of centered cocycles is the $L^1(\mathbb{P})$ closure of gradients $F(\omega, x, y) = \varphi(T_y\omega) - \varphi(T_x\omega)$, see [28, Lemma C.3]. If B is a stationary L^1 cocycle, there exists a unique vector $h(B) \in \mathbb{R}^2$ such that

$$\mathbb{E}[B(0, e_i)] = -h(B) \cdot e_i \qquad \text{for } i \in \{1, 2\}. \tag{3.2}$$

Then

$$F(x, y) = h(B) \cdot (x - y) - B(x, y) \tag{3.3}$$

is a centered cocycle.

Theorem 3.2. *In the point-to-line case, the limits in* (2.9) *have these variational representations: for* $0 < \beta < \infty$

$$g_{pl}^{\beta}(h) = \inf_{F \in \mathcal{C}_0} \mathbb{P}\text{-}\operatorname*{ess\,sup}_{\omega} \ \beta^{-1} \log \sum_{i \in \{1,2\}} \tfrac{1}{2} e^{\beta V(\omega_0) + \beta h \cdot e_i + \beta F(\omega, 0, e_i)} \tag{3.4}$$

and for $\beta = \infty$

$$g_{pl}^{\infty}(h) = \inf_{F \in \mathcal{C}_0} \mathbb{P}\text{-}\operatorname*{ess\,sup}_{\omega} \max_{i \in \{1,2\}} \{V(\omega_0) + h \cdot e_i + F(\omega, 0, e_i)\}. \tag{3.5}$$

Furthermore, in both cases a minimizing cocycle $F \in \mathcal{C}_0$ *exists.*

The existence of a minimizing cocycle is proved by taking a weak limit of averages of increments of $G_{(N)}(h)$.

There is a duality between the point-to-point and point-to-line limits, given by

$$g_{\text{pp}}^{\beta}(\xi) = \inf_{h \in \mathbb{R}^2} \{g_{\text{pl}}^{\beta}(h) - h \cdot \xi\} \qquad \text{for } \xi \in \mathcal{U}^{\circ} \tag{3.6}$$

and

$$g_{\text{pl}}^{\beta}(h) = \sup_{\xi \in \mathcal{U}^{\circ}} \{g_{\text{pp}}^{\beta}(\xi) + h \cdot \xi\} \qquad \text{for } h \in \mathbb{R}^2. \tag{3.7}$$

Let us say that $\xi \in \mathcal{U}^{\circ}$ and $h \in \mathbb{R}^2$ are dual if

$$g_{\text{pp}}^{\beta}(\xi) = g_{\text{pl}}^{\beta}(h) - h \cdot \xi. \tag{3.8}$$

Formulas (3.4)–(3.5) and the duality (3.6) lead to analogous variational formulas for the point-to-point case.

Theorem 3.3. *In the point-to-point case, the limits in* (2.8) *have these variational formulas for* $\xi \in \mathcal{U}^{\circ}$. *For* $0 < \beta < \infty$

$$g_{pp}^{\beta}(\xi) \ = \ \inf_{B \in \mathcal{C}} \mathbb{P}\text{-}\operatorname*{ess\,sup}_{\omega} \ \beta^{-1} \log \sum_{i \in \{1,2\}} \tfrac{1}{2} e^{\beta V(\omega_0) - \beta B(\omega, 0, e_i) - \beta h(B) \cdot \xi} \tag{3.9}$$

and for $\beta = \infty$

$$g_{pp}^{\infty}(\xi) \;=\; \inf_{B\in\mathcal{C}} \; \mathbb{P}\text{-}\operatorname*{ess\,sup}_{\omega} \; \max_{i\in\{1,2\}} \{V(\omega_0) - B(\omega, 0, e_i) - h(B)\cdot\xi\} \tag{3.10}$$

where the infima are over stationary L^1 *cocyles* B*. In each case a minimizing cocycle* B *exists.*

The variational formulas presented in this section were proved in articles [12, 27, 28], first for $\beta < \infty$ by developing a suitable large deviation theory, and then for last-passage percolation by letting $\beta \to \infty$.

4. Busemann functions for the log-gamma polymer

To illustrate cocycles that solve the variational formulas, we turn to the exactly solvable *log-gamma polymer*. For this model we can prove that so-called *Busemann functions* exist. These are limits of gradients of the free energy. In certain exactly solvable models Busemann functions have tractable probability distributions. Then we can realize the program alluded to in the introduction, namely derivation of KPZ fluctuation exponents through control of the fluctuations of Busemann functions. This is the topic of Section 5 below.

Fix $0 < \rho < \infty$ and let weights $\{\omega_x\}$ be i.i.d. Gamma(ρ) distributed. The partition functions of the log-gamma polymer are defined by

$$Z_{u,v} = \sum_{x_{\cdot}\in\Pi_{u,v}} \prod_{i=0}^{|v-u|_1-1} \omega_{x_i}^{-1} \qquad \text{for } u \le v \text{ in } \mathbb{Z}^2. \tag{4.1}$$

Note that the weight at u is included and v excluded. This corresponds to (2.2) with

$$V(\omega_x) = -\log\omega_x + \log 2. \tag{4.2}$$

The inverse temperature parameter β is not explicitly present in the log-gamma polymer so it is fixed at $\beta = 1$ and dropped from the notation. In a sense, parameter ρ plays the role of temperature.

In addition to the duality (3.8) between tilts $h \in \mathbb{R}^2$ and directions $\xi \in \mathcal{U}^\circ$, a third variable $\lambda \in (0, \rho)$ is in bijective correspondence with directions $\xi = (s, 1-s) \in \mathcal{U}^\circ$ via the equation

$$s\Psi_1(\lambda) - (1-s)\Psi_1(\rho-\lambda) = 0. \tag{4.3}$$

We call ξ the *characteristic direction* of (λ, ρ).

Theorem 4.1. *Fix* $\xi \in \mathcal{U}^\circ$ *and let* $\lambda = \lambda(\xi) \in (0,\rho)$ *be determined by* (4.3)*. Then on* $(\Omega, \mathcal{S}, \mathbb{P})$ *there exists a stationary* L^1 *cocycle* $\{B^\xi(x,y) : x, y \in \mathbb{Z}^2\}$ *with the following properties.*

(a) *Suppose a point* $v \in \mathbb{N}^2$ *tends to infinity in the first quadrant so that* $v/|v|_1 \to \xi$*. Then for all* $x, y \in \mathbb{Z}^2$ *these almost sure limits hold:*

$$B^\xi(x,y) = \lim_{v\to\infty} \bigl(\log Z_{x,v} - \log Z_{y,v}\bigr). \tag{4.4}$$

(b) *Let $h = (h_1, h_2)$ be dual to ξ in the sense of* (3.8)*. Then for $i \in \{1, 2\}$*

$$\lim_{N\to\infty} \bigl(\log Z_{(N)}(h) \circ T_x - \log Z_{(N-1)}(h) \circ T_{x+e_i}\bigr) = B^\xi(x, x + e_i) + h_i. \quad (4.5)$$

(c) *For each $x \in \mathbb{Z}^2$ the process $\{e^{-B^\xi(x+ie_1, x+(i+1)e_1)} : i \in \mathbb{Z}_+\}$ is i.i.d. Gamma(λ), the process $\{e^{-B^\xi(x+je_2, x+(j+1)e_2)} : j \in \mathbb{Z}_+\}$ is i.i.d. Gamma($\rho - \lambda$), and these two processes are independent of each other.*

Parts (a) and (c) of the theorem above are contained in Theorem 4.1 of [13] and part (b) is in Theorem 6.1 of the same paper. From part (c) above we can calculate the mean vector $h(B^\xi)$ defined in (3.2):

$$h(B^\xi) = -\bigl(\mathbb{E}[B^\xi(0, e_1)]\,,\, \mathbb{E}[B^\xi(0, e_2)]\bigr) = \bigl(\Psi_0(\lambda), \Psi_0(\rho - \lambda)\bigr) \quad (4.6)$$

again with $\lambda = \lambda(\xi)$ defined by (4.3).

The next theorem gives the minimizers of the variational formulas. Inverse temperature is fixed at $\beta = 1$ and dropped from the notation.

Theorem 4.2. *Fix $\xi = (s, 1 - s) \in \mathcal{U}^\circ$ and let $\lambda = \lambda(\xi) \in (0, \rho)$ be determined by* (4.3)*. Let B^ξ be the cocycle in Theorem 4.1.*

(a) *$h = (h_1, h_2) \in \mathbb{R}^2$ is dual to $\xi \in \mathcal{U}^\circ$ if and only if*

$$h_1 - h_2 = \Psi_0(\lambda) - \Psi_0(\rho - \lambda). \quad (4.7)$$

(b) *B^ξ is a minimizer in* (3.9) *and the essential supremum disappears: for $\mathbb{P}$-a.e. ω,*

$$g_{pp}(\xi) \;=\; \log \sum_{i\in\{1,2\}} \tfrac{1}{2} e^{V(\omega_0) - B^\xi(\omega, 0, e_i) - h(B^\xi)\cdot\xi} \;=\; -\xi \cdot h(B^\xi). \quad (4.8)$$

(c) *Suppose $h = h(B^\xi) + (t, t)$ for some $t \in \mathbb{R}$. Then h and ξ are dual. Cocycle $F(\omega, x, y) = h(B^\xi) \cdot (x - y) - B^\xi(\omega, x, y)$ is a minimizer in* (3.4) *for this h and the essential supremum disappears: for $\mathbb{P}$-a.e. ω and $j \in \{1, 2\}$,*

$$g_{pl}(h) = \log \sum_{i\in\{1,2\}} \tfrac{1}{2} e^{V(\omega_0) + h\cdot e_i + F(\omega, 0, e_i)} = (h - h(B^\xi)) \cdot e_j = t. \quad (4.9)$$

The theorem above is collected from results in Section 5 of [13].

Remark 4.3. The organization of this section does not represent the chronological order of discovery. The formula

$$\begin{aligned} g_{\text{pp}}(\xi) &= -\xi \cdot \bigl(\Psi_0(\lambda(\xi)), \Psi_0(\rho - \lambda(\xi))\bigr) \\ &= \inf_{\nu\in(0,\rho)} \xi \cdot \bigl(-\Psi_0(\nu), -\Psi_0(\rho - \nu)\bigr) \end{aligned} \quad (4.10)$$

was computed first in [32] and then in [14] before the Busemann functions were derived.

An additional corollary of the Busemann limits of Theorem 4.1 is the convergence of the quenched polymer measures to random walks in correlated random environments. Define a transition probability on $\mathbb{Z}^2$ by

$$\pi^\xi(\omega, x, x+e_i) = \frac{e^{-B^\xi(\omega,x,x+e_i)}}{e^{-B^\xi(\omega,x,x+e_1)} + e^{-B^\xi(\omega,x,x+e_2)}}, \qquad i \in \{1,2\},$$

and let $P^{\omega,\xi}$ be the path measure of the Markov chain that starts at 0 and follows transitions $\pi^\xi(\omega, x, x+e_i)$.

Corollary 4.4. *Consider the log-gamma polymer with i.i.d. Gamma(ρ) weights $\{\omega_x\}$, V as in* (4.2)*, and $\beta = 1$. Let $Q_{0,v}$ and $Q_{(N)}$ be the quenched polymer measures defined by* (2.1) *and* (2.3)*. Fix $\xi \in \mathcal{U}^\circ$ and let h be dual to ξ. Let $v \in \mathbb{N}^2$ tend to infinity in the first quadrant so that $v/|v|_1 \to \xi$. Let $N \to \infty$. Then for $\mathbb{P}$-a.e. ω both measures $Q_{0,v}$ and $Q_{(N)}$ converge weakly to $P^{\omega,\xi}$.*

Proof. Consider the point-to-point case. Fix an admissible path segment $(x_k)_{k=0}^m$.

$$\begin{aligned} &Q_{0,v}\{X_k = x_k \text{ for } k = 0,\dots,m\} = \frac{Z_{x_m,v}}{Z_{0,v}} 2^{-m} \exp\Big(\sum_{k=0}^{m-1} V(\omega_{x_k}) \Big) \\ &\longrightarrow e^{-B^\xi(0,x_m)} 2^{-m} \exp\Big(\sum_{k=0}^{m-1} V(\omega_{x_k}) \Big) = \prod_{k=0}^{m-1} \frac{e^{-B^\xi(x_k,x_{k+1})}}{\omega_{x_k}} \\ &= P^{\omega,\xi}\{X_k = x_k \text{ for } k = 0,\dots,m\}. \end{aligned}$$

The last step relied on $\omega_x = e^{-B^\xi(\omega,x,x+e_1)} + e^{-B^\xi(\omega,x,x+e_2)}$ which comes from the limit in Theorem 4.1(a). □

The weak limit above is in Theorem 7.1 of [13]. Theorem 8.2 of that paper exhibits a stationary, ergodic invariant distribution for the environment as seen from the particle for the random walk in random environment with transition π^ξ. Whether the environment process initialized as in Corollary 4.4 converges to the stationary process is currently open.

5. Stationary log-gamma polymer and fluctuation exponents

The Busemann limits of Theorem 4.1 represent a stationary version of the polymer process $\{Z_{u,v}\}_{u \le v}$ of (4.1), as explained next. Continue with a fixed $\xi \in \mathcal{U}^\circ$, $\lambda = \lambda(\xi) \in (0, \rho)$ determined by (4.3), and the cocycle B^ξ of Theorem 4.1. For $x \in \mathbb{Z}^2$ define weight

$$\begin{aligned} \tau_{x-e_i,x} &= e^{-B^\xi(\omega,x-e_i,x)} \quad \text{on the edge } \{x-e_i, x\} \text{ for } i \in \{1,2\}, \text{ and} \\ \text{weight} \quad \sigma_x &= e^{-B^\xi(\omega,x-e_1,x)} + e^{-B^\xi(\omega,x-e_2,x)} \quad \text{on the vertex } x. \end{aligned}$$

Theorem 5.1. *For any $x \in \mathbb{Z}^2$, the weights*

$$\{\tau_{x+(i-1)e_1,x+ie_1},\ \tau_{x+(j-1)e_2,x+je_2},\ \sigma_{x+(i,j)} : i,j \in \mathbb{N}\} \tag{5.1}$$

are independent with marginal distributions

$$\begin{aligned} &\tau_{z-e_1,z} \sim \textit{Gamma}(\lambda), \quad \tau_{z-e_2,z} \sim \textit{Gamma}(\rho-\lambda), \\ &\textit{and} \quad \sigma_z \sim \textit{Gamma}(\rho). \end{aligned} \tag{5.2}$$

The collection in (5.1) should be viewed as consisting of boundary edge weights τ on the axes and bulk weights σ in the first quadrant relative to the origin at x. This is a stationary log-gamma polymer process for partition functions that combine boundary weights and bulk weights. If we place the origin at 0 then such a partition function is defined by

$$Z_{0,v}^{(\lambda)} = \sum_{x_\cdot \in \Pi_{0,v}} \Big(\prod_{i=1}^{t_{\text{exit}}} \tau_{x_{i-1},\, x_i}^{-1} \Big) \Big(\prod_{j=t_{\text{exit}}+1}^{|v|_1} \sigma_{x_j}^{-1} \Big), \quad v \in \mathbb{Z}_+^2, \tag{5.3}$$

where the exit time t_{exit} of the path $(x_k)_{k\geq 0}$ from the boundary is $t_{\text{exit}} = t_{e_1} \vee t_{e_2}$ with

$$t_{e_1} = \max\{k \geq 0 : x_i = (i,0) \text{ for } 0 \leq i \leq k\} \tag{5.4}$$

and

$$t_{e_2} = \max\{\ell \geq 0 : x_j = (0,j) \text{ for } 0 \leq j \leq \ell\}. \tag{5.5}$$

For each path $t_{e_1} \wedge t_{e_2} = 0$. The superscript (λ) on $Z_{0,v}^{(\lambda)}$ denotes the stationary process with parameter λ. Then one shows inductively that

$$\tau_{x-e_i,x} = \frac{Z_{0,x-e_i}^{(\lambda)}}{Z_{0,x}^{(\lambda)}} \qquad \forall x \in \mathbb{Z}_+^2 \text{ and } i \in \{1,2\} \text{ such that } x - e_i \in \mathbb{Z}_+^2.$$

Theorem 5.1 tells us that the joint distributions of the ratios of partition functions $Z_{0,v}^{(\lambda)}$ are invariant under lattice translations. This is the precise sense in which the process $\{Z_{0,v}^{(\lambda)}\}_{v\in\mathbb{Z}_+^2}$ is stationary.

The partition function (5.3) of the stationary polymer can be also written as (with $v = (m,n) \in \mathbb{N}^2$)

$$\begin{aligned} Z_{0,v}^{(\lambda)} \;=\; & \sigma_v^{-1} \sum_{k=1}^{m} \Big(\prod_{i=1}^{k} \tau_{(i-1)e_1,\, ie_1}^{-1} \Big) Z_{(k,1),v} \\ & + \; \sigma_v^{-1} \sum_{\ell=1}^{n} \Big(\prod_{j=1}^{\ell} \tau_{(j-1)e_2,\, je_2}^{-1} \Big) Z_{(1,\ell),v} \end{aligned} \tag{5.6}$$

where $Z_{u,v}$ denotes the original log-gamma partition function (4.1) but in terms of the i.i.d. Gamma(ρ) weights $\{\sigma_x\}$. This gives a *coupling* (simultaneous construction) of the stationary polymer process and the original polymer process. Through this coupling the stationary process can be used to study the original log-gamma polymer.

We indicate some of the benefits of the stationary log-gamma process. The characteristic direction is relevant again. The next theorem states the KPZ fluctuation exponent $1/3$ for the free energy of the stationary log-gamma polymer.

Theorem 5.2 (Theorem 2.1 [32]). *Let $\lambda \in (0,\rho)$ and $\xi \in \mathcal{U}^\circ$ correspond to each other via* (4.3)*. Let $Z_{0,v}^{(\lambda)}$ be the stationary log-gamma partition function defined by* (5.3) *with weights distributed as in* (5.2)*. Then there exists a constant $0 < C < \infty$ such that, for $N \geq 1$,*

$$C^{-1} N^{2/3} \leq \mathbb{V}\text{ar}(\log Z_{0,\lfloor N\xi \rfloor}^{(\lambda)}) \leq C N^{2/3}.$$

If ξ is not the characteristic direction for (λ, ρ) then asymptotically the fluctuations of $\log Z^{(\lambda)}_{0,\lfloor N\xi \rfloor}$ are Gaussian of order $N^{-1/2}$. In other words, the classical central limit theorem rules. This is because the Gaussian fluctuations of the boundary weights τ in (5.6) dominate the fluctuations of the partition functions Z.

The coupling (5.6) allows us to pass fluctuation bounds to the original log-gamma polymer with i.i.d. weights. However, the result is weaker than the one in Theorem 5.2.

Theorem 5.3. *Let $0 < \rho < \infty$ and let $Z_{u,v}$ denote the partition function of* (4.1) *with i.i.d. Gamma(ρ) weights. Let $g_{pp}(\xi)$ be the limiting free energy density from* (4.10)*. Then for $1 \le p < 3/2$ there exists a constant $0 < C_p < \infty$ such that, for $N \ge 1$,*

$$C_p^{-1} N^{p/3} \le \mathbb{E}\big[\, |\log Z_{0,\lfloor N\xi \rfloor} - N g_{pp}(\xi)\,|^p \big] \le C_p N^{p/3}.$$

The upper bound is in Theorem 2.4 of [32]. The lower bound is as yet unpublished, although we have published the analogous result for the O'Connell-Yor semidiscrete polymer in [24]. The theorem is proved by coupling $Z_{0,\lfloor N\xi \rfloor}$ with a stationary polymer $Z^{(\lambda)}_{0,\lfloor N\xi \rfloor}$ with parameter λ chosen according to (4.3). This brings the quantities $\log Z^{(\lambda)}_{0,\lfloor N\xi \rfloor}$ and $\log Z_{0,\lfloor N\xi \rfloor}$ within $O(N^{1/3})$ of each other. Details appear in [32].

Results on Busemann functions and fluctuation exponents, analogous to those in Sections 4 and 5 for the log-gamma polymer, are valid also for the exactly solvable last-passage percolation processes, where the weights $\{\omega_x\}$ are i.i.d. exponential or geometric random variables. Results on fluctuation exponents were derived in [4]. Extension of these properties beyond exactly solvable models, to the corner growth model (2.6) with general weights, is currently under way.

Acknowledgements. T. Seppäläinen was partially supported by National Science Foundation grant DMS-1306777 and by the Wisconsin Alumni Research Foundation.

References

[1] D. Aldous and P. Diaconis, *Hammersley's interacting particle process and longest increasing subsequences*, Probab. Theory Related Fields **103**(2) (1995), 199–213.

[2] Jinho Baik, Percy Deift, and Kurt Johansson, *On the distribution of the length of the longest increasing subsequence of random permutations*, J. Amer. Math. Soc. **12**(4) (1999), 1119–1178.

[3] Yuri Bakhtin, Eric Cator, and Konstantin Khanin, *Space-time stationary solutions for the Burgers equation*, J. Amer. Math. Soc. **27**(1) (2014), 193–238.

[4] Márton Balázs, Eric Cator, and Timo Seppäläinen, *Cube root fluctuations for the corner growth model associated to the exclusion process*, Electron. J. Probab. **11** (2006), no. 42, 1094–1132 (electronic).

[5] Alexei Borodin and Ivan Corwin, *Macdonald processes*, Probab. Theory Related Fields **158**(1-2) (2014), 225–400.

[6] Alexei Borodin, Ivan Corwin, and Daniel Remenik, *Log-gamma polymer free energy fluctuations via a Fredholm determinant identity*, Comm. Math. Phys. **324**(1) (2013), 215–232.

[7] Eric Cator and Leandro P. R. Pimentel, *Busemann functions and equilibrium measures in last passage percolation models*, Probab. Theory Related Fields **154**(1-2) (2012), 89–125.

[8] Henry Cohn, Noam Elkies, and James Propp, *Local statistics for random domino tilings of the Aztec diamond*, Duke Math. J. **85**(1) (1996), 117–166.

[9] Ivan Corwin, *The Kardar-Parisi-Zhang equation and universality class*, Random Matrices Theory Appl. **1**(1) (2012), 1130001, 76.

[10] J. Theodore Cox and Richard Durrett, *Some limit theorems for percolation processes with necessary and sufficient conditions*, Ann. Probab. **9**(4) (1981), 583–603.

[11] Michael Damron and Jack Hanson, *Busemann functions and infinite geodesics in two-dimensional first-passage percolation*, Comm. Math. Phys. **325**(3) (2014), 917–963.

[12] Nicos Georgiou, Firas Rassoul-Agha, and Timo Seppäläinen, *Variational formulas and cocycle solutions for directed polymer and percolation models*, arXiv:1311.3016, 2013.

[13] Nicos Georgiou, Firas Rassoul-Agha, Timo Seppäläinen, and Atilla Yilmaz, *Ratios of partition functions for the log-gamma polymer*, Ann. Probab. (to appear) arXiv:1303.1229, 2013.

[14] Nicos Georgiou and Timo Seppäläinen, *Large deviation rate functions for the partition function in a log-gamma distributed random potential*, Ann. Probab. **41**(6) (2013), 4248–4286.

[15] J. M. Hammersley and D. J. A. Welsh, *First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory*, In Proc. Internat. Res. Semin., Statist. Lab., Univ. California, Berkeley, Calif, pp. 61–110. Springer-Verlag, New York, 1965.

[16] John M. Hammersley, *A few seedlings of research*, In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics, pp. 345–394, Berkeley, Calif., 1972. Univ. California Press.

[17] Christopher Hoffman, *Coexistence for Richardson type competing spatial growth models*, Ann. Appl. Probab. **15**(1B) (2005), 739–747.

[18] D. A. Huse and C. L. Henley, *Pinning and roughening of domain wall in Ising systems due to random impurities*, Phys. Rev. Lett. **54** (1985), 2708–2711.

[19] William Jockusch, James Propp, and Peter Shor, *Random domino tilings and the arctic circle theorem*, arXiv:math/9801068.

[20] Kurt Johansson, *Shape fluctuations and random matrices*, Comm. Math. Phys. **209**(2) (2000), 437–476.

[21] K. Kardar, G. Parisi, and Y. Zhang, *Dynamic scaling of growing interfaces*, Phys. Rev. Lett. **56** (1986), 889–892.

[22] Arjun Krishnan, *Variational formula for the time-constant of first-passage percolation I: Homogenization*, 2013. arXiv:1311.0316.

[23] J. B. Martin, *Last-passage percolation with general weight distribution*, Markov Process. Related Fields **12**(2) (2006), 273–299.

[24] Gregorio Moreno Flores, Timo Seppäläinen, and Benedek Valkó, *Fluctuation exponents for directed polymers in the intermediate disorder regime*, 2013. arXiv:1312.0519.

[25] Charles M. Newman, *A surface view of first-passage percolation*, In Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), pp. 1017–1023, Basel, 1995. Birkhäuser.

[26] Jeremy Quastel, *Weakly asymmetric exclusion and KPZ*, In Proceedings of the International Congress of Mathematicians. Volume IV, pp. 2310–2324, New Delhi, 2010. Hindustan Book Agency.

[27] Firas Rassoul-Agha and Timo Seppäläinen, *Quenched point-to-point free energy for random walks in random potentials*, Probab. Theory Related Fields **158**(3-4) (2014), 711–750.

[28] Firas Rassoul-Agha, Timo Seppäläinen, and Atilla Yilmaz, *Quenched free energy and large deviations for random walks in random potentials*, Comm. Pure Appl. Math. **66**(2) (2013), 202–244.

[29] Hermann Rost, *Nonequilibrium behaviour of a many particle process: density profile and local equilibria*, Z. Wahrsch. Verw. Gebiete **58**(1) (1981), 41–53.

[30] Timo Seppäläinen, *A microscopic model for the Burgers equation and longest increasing subsequences*, Electron. J. Probab. **1** (1996), no. 5, approx. 51 pp. (electronic).

[31] ———, *Hydrodynamic scaling, convex duality and asymptotic shapes of growth models*, Markov Process. Related Fields **4**(1) (1998), 1–26.

[32] ———, *Scaling for a one-dimensional directed polymer with boundary conditions*, Ann. Probab. **40**(1) (2012), 19–73.

[33] Herbert Spohn, *Stochastic integrability and the KPZ equation*, arXiv:1204.2657.

[34] Craig A. Tracy and Harold Widom, *Distribution functions for largest eigenvalues and their applications*, In Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002), pp. 587–596, Beijing, 2002. Higher Ed. Press.

University of Wisconsin-Madison, Mathematics Department, Van Vleck Hall, 480 Lincoln Dr., Madison WI 53706-1388, USA

E-mail: seppalai@math.wisc.edu

Criticality and Phase Transitions: five favorite pieces

Vladas Sidoravicius

Abstract. We present few recent results concerning the phase transition and behavior of classical equilibrium and non-equilibrium systems at criticality. Five topics are discussed: a) continuity of the phase transition for Bernoulli percolation, Ising and Potts models; b) geometry of critical percolation clusters in the context of self-destructive percolation; c) non-equilibrium phase transitions and critical behavior of conservative lattice gasses; d) dynamic phase transitions for KPZ growth models and solution of slow bond problem; e) solution of Coffman-Gilbert conjecture.

Mathematics Subject Classification (2010). Primary 60K35; Secondary 82B30.

Keywords. phase transition, criticality, percolation, Ising model, Potts model, non-equilibrium phase transition, dynamic phase transition, greedy algorithm.

1. Introduction

Equilibrium and non-equilibrium Statistical Mechanics offers many classes of models which exhibit non-trivial behavior and undergo different types of phase transitions. While equilibrium systems are relatively well understood away from criticality, and, in some cases, at criticality too, understanding of their critical behavior below upper critical dimension poses substantial mathematical difficulties and still remains poorly understood. For non-equilibrium systems situation is much less clear even away from criticality. However in the last several years some progress has been made in various directions, bringing solutions to some old conjectures, developing new mathematical concepts and tools, and opening new areas of research. Constantly growing "global" understanding of critical systems on the theoretical and rigorous levels led in some cases to establishment of new fruitful links and connections between different subfields, such as models of self-organized criticality and usual non-equilibrium critical systems, models exhibiting dynamic phase transition and classical interacting particle systems.

This note contains five separte parts presenting recent progress and discussing some open problems in following selected directions: a) classical question of continuity of the phase transition for Bernoulli percolation, Ising and Potts models; b) geometry of critical percolation clusters in the context of self-destructive percolation and its connections to "ghost" forest fire models; c) non-equilibrium phase transitions (absorbing state phase transitions) and critical behavior of conservative lattice gasses; d) criticality for dynamic phase transitions, KPZ type growth systems in presence of a columnar defect and the solution of slow bond problem; e) solution of Coffman-Gilbert conjecture.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

An interested reader is referred to an extended version of this article [105].

2. Continuity of the phase transition

Determining whether a phase transition is continuous or discontinuous is one of the fundamental questions in statistical physics. Bernoulli percolation has offered the mathematicians a setup to develop techniques to prove either continuity or discontinuity of the phase transition, which in the case of continuity corresponds to the absence of an infinite cluster at criticality. Harris [66] proved that the nearest neighbor bond percolation model with parameter $1/2$ on $\mathbb{Z}^2$ does not contain an infinite cluster almost surely. Viewed together with Kesten's result that $p_c \leq 1/2$ [72], it provided the first proof of such type of statement. Since the original proof of Harris, a few alternative arguments have been found for planar graphs. In the late eighties, dynamic renormalization ideas were successfully applied to prove continuity in octants and half spaces of $\mathbb{Z}^d$, $d \geq 3$, [14, 15]. The continuity was also proved for $\mathbb{Z}^d$ with $d \geq 19$ using the lace expansion technique [65], and for non-amenable Cayley graphs using mass-transport arguments [24]. Despite all these developments, a general argument to prove the continuity of the phase transition for the nearest neighbor Bernoulli percolation on arbitrary lattices is still missing, and the fact that the Bernoulli percolation undergoes a continuous phase transition on $\mathbb{Z}^3$ still represents one of the major open questions in the field.

Bernoulli percolation in slabs. Here we state the continuity for Bernoulli percolation on a class of non-planar lattices, namely slabs.

Consider the graph $\mathbb{S}$, called *slab* of width k, given by the vertex set $\mathbb{Z}^2 \times \{0, \dots, k\}$ and edges between nearest neighbors. In what follows, $\mathbf{P}_p$ denotes the Bernoulli bond percolation measure with parameter p on $\mathbb{S}$ defined as follows: every edge of $\mathbb{Z}^2 \times \{0, \dots, k\}$ is *open* with probability p (if it is not open, it is said to be *closed*) independently of the other edges. Let $p_c(k)$ be the critical parameter of Bernoulli percolation on $\mathbb{S}$. Let B be a subset of $\mathbb{Z}^3$, the event $\{0 \overset{B}{\longleftrightarrow} \infty\}$ denotes the existence of an infinite path of open edges in B starting from 0.

Theorem 2.1 ([45]). *For any $k > 0$, $\mathbf{P}_{p_c(k)}[0 \overset{\mathbb{S}}{\longleftrightarrow} \infty] = 0$.*

For *site* percolation on $\mathbb{S}_2$, an *ad hoc* argument was provided in [38]. Proof of [45] works equally well (with suitable modifications) for any graph of the form $\mathbb{Z}^2 \times G$, where G is finite. This includes $G = \{0, \dots, k\}^{d-2}$ for $d \geq 3$.

Similarly, symmetric finite range percolation on $\mathbb{Z}^2$ can be treated via the same techniques (once again, relevant modifications must be done). Let us state the result in this setting. Let $\mathbf{p} \in [0,1]^{\mathbb{Z}^2}$ be a set of edge-weight parameters, and $M > 0$. We consider functions $\mathbf{p}$'s that are M-supported (meaning $\mathbf{p}_z = 0$ for $|z| \geq M$) and invariant under reflection and $\pi/2$-rotation (meaning that for all z, $\mathbf{p}_{\mathrm{i}z} = \mathbf{p}_{\bar{z}} = \mathbf{p}_z$). Consider the graph with vertex set $\mathbb{Z}^2$ and edges between any two vertices and the percolation $\mathbf{P}_{\mathbf{p}}$ defined as follows: the edge (x, y) is open with probability $\mathbf{p}_{x-y}$, independently of the other edges.

Theorem 2.2 ([45]). *$M > 0$. The probability $\mathbf{P}_{\mathbf{p}}[0 \longleftrightarrow \infty]$ is continuous, when viewed as a function defined on the set of M-supported and invariant $\mathbf{p}$'s.*

From the slab to $\mathbb{Z}^3$? The fact that $\mathbb{Z}^2 \times \{0, \dots, k\}^{d-2}$ is approximating $\mathbb{Z}^d$ when k tends

to infinity suggests that the non-percolation on slabs could shed a new light on the problem of proving the absence of infinite cluster (almost surely) for critical percolation on $\mathbb{Z}^d$. Nevertheless, we wish to highlight that this is not immediate. Indeed, while $p_c(k)$ is known to converge to $p_c(\mathbb{Z}^3)$ [63], passing at the limit requires a new ingredient. For instance, a uniform control (in k) on the explosion of the infinite-cluster density for p tending to the critical point would be sufficient.

Proposition 2.3. *Let $f : [0,1] \to \mathbb{R}$ be a continuous function such that $f(0) = 0$. If for any $k \geq 0$ and any $p \in (0,1)$,*

$$\mathbf{P}_p[0 \overset{\mathbb{S}}{\longleftrightarrow} \infty] \leq f(p - p_c(k)),$$

then $\mathbf{P}_{p_c(\mathbb{Z}^3)}[0 \overset{\mathbb{Z}^3}{\longleftrightarrow} \infty] = 0$.

It is natural to expect that proving the existence of f is roughly of the same difficulty as attacking the problem directly on $\mathbb{Z}^3$. Nevertheless, it could be that a suitable renormalization argument enables one to prove the existence of f.

Continuity of phase transition for Ising model. Here we present an answer to a question of the continuity at the critical point of the spontaneous magnetization of the standard three dimensional Ising model. More generally, the model may be formulated on a graph, whose vertex set and edge set we denote by $\mathbb{G}$ and $E \subset \mathbb{G}^2$ correspondingly. Associated with the sites are ± 1 valued spin variables, whose configuration is denoted $\sigma = (\sigma_x : x \in \mathbb{G})$.

For a general ferromagnetic pair interaction, the system's Hamiltonian defined for finite subsets $\Lambda \subset \mathbb{G}$ and boundary conditions $\tau \in \{-1,0,1\}^{\mathbb{G}\setminus\Lambda}$ is given by the function

$$H^\tau_\Lambda(\sigma) \;:=\; -\sum_{x\in\Lambda} h\sigma_x \;-\; \sum_{\{x,y\}\subset\Lambda : x\neq y} J_{x,y}\sigma_x\sigma_y \;-\; \sum_{x\in\Lambda : y\in\mathbb{G}\setminus\Lambda} J_{x,y}\sigma_x\tau_y\,, \tag{2.1}$$

for any $\sigma \in \{-1,1\}^\Lambda$, where $(J_{x,y})_{x,y\in\mathbb{Z}^d}$ is a family of nonnegative *coupling constants*, and h is the magnetic field.

For $\beta \in (0,\infty)$, finite volume Gibbs states with boundary conditions τ are given by probability measures on the spaces of configurations in finite subsets $\Lambda \subset \mathbb{G}$ under which the expected values of functions $f : \{-1,1\}^\Lambda \to \mathbb{R}$ are

$$\langle f\rangle^\tau_{\Lambda,\beta,h} = \sum_{\sigma\in\{-1,1\}^\Lambda} f(\sigma)\frac{e^{-\beta H^\tau_\Lambda(\sigma)}}{Z^\tau(\Lambda,\beta,h)}\,,$$

where the sum is normalized by the partition function $Z^\tau(\Lambda,\beta,h)$ so that $\langle 1\rangle^\tau_{\Lambda,\beta,h} = 1$. Of particular interest is the following pair of boundary conditions (b.c.):

free b.c.: $\tau_x = 0$ for all $x \in \mathbb{G}\setminus\Lambda$.

plus b.c.: $\tau_x = 1$ for all $x \in \mathbb{G}\setminus\Lambda$.

The corresponding measures are denoted by $\langle\cdot\rangle^0_{\Lambda,\beta,h}$ and $\langle\cdot\rangle^+_{\Lambda,\beta,h}$ respectively.

For each of these two boundary conditions, the finite volume Gibbs states are known to converge to the corresponding infinite-volume Gibbs measures. We focus here on the case $\mathbb{G} = \mathbb{Z}^d$, and interactions which are:

c1) translation invariant: $J_{x,y} = J_{0,y-x}$,

c2) ferromagnetic: $J_{x,y} \geq 0$,

c3) locally finite: $|J| := \sum_{x\in\mathbb{Z}^d} J_{0,x} < \infty$.

c4) aperiodic: for any $x \in \mathbb{Z}^d$, there exist $0 = x_0, x_1, \ldots, x_{m-1}, x_m = x$ such that $J_{x_0,x_1} J_{x_1,x_2}, \ldots, J_{x_{m-1},x_m} > 0$.

Of particular interest is the model's phase transition, which in the (β, h) plane occurs along the $h = 0$ line and is reflected in the nonvanishing of the symmetry breaking order parameter:

$$m^*(\beta) \;:=\; \langle\sigma_0\rangle^+_\beta \,. \tag{2.2}$$

For temperatures ($T \equiv \beta^{-1}$) at which $m^*(\beta) > 0$, the mean magnetization at nonzero magnetic field h changes discontinuously at $h = 0$. The discontinuity is symptomatic of the co-existence of two distinct Gibbs equilibrium states :

$$\langle\cdot\rangle^+_\beta = \lim_{h\searrow 0} \langle\cdot\rangle_{\beta,h} \qquad \langle\cdot\rangle^-_\beta = \lim_{h\nearrow 0} \langle\cdot\rangle_{\beta,h} \tag{2.3}$$

which carry the residual magnetizations:

$$\langle\sigma_0\rangle^\pm_\beta \;=\; \pm m^*(\beta)\,, \tag{2.4}$$

with $m^*(\beta)$ customarily referred to as the *spontaneous magnetization*.

Property c3) guarantees that at small β (in particular, for $\beta < |J|^{-1}$, see e.g. [50, 56]) $m^*(\beta) = 0$, and there is no symmetry breaking. However, in dimensions $d > 1$ each such model exhibits a phase transition at some $\beta_c \in (|J|^{-1}, \infty)$, with $m^*(\beta) > 0$ for $\beta > \beta_c$ [89]. For $d = 1$ such a transition occurs if $J_{x-y} \geq 1/|x-y|^\alpha$ with $\alpha \in (1, 2)$ [51] and also for the boundary value $\alpha = 2$ [5], in which case $m^*(\beta)$ is discontinuous at β_c [5, 108]).

Relevant to the continuity of the spontaneous magnetization is the Long Range Order (LRO) parameter:

$$M_{LRO}(\beta) \;:=\; \lim_{n\to\infty} \frac{1}{|\Lambda_n|} \sum_{x\in\Lambda_n} \langle\sigma_0\sigma_x\rangle^0_\beta \,, \tag{2.5}$$

where $\Lambda_n = [-n, n]^d$, for the model on $\mathbb{Z}^d$ (the limit existing by monotonicity arguments), or the LRO parameter's variant

$$\widetilde{M}_{LRO}(\beta) \;:=\; \inf_{B\subset\mathbb{Z}^d, |B|<\infty} \frac{1}{|B|^2} \sum_{x,y\in B} \langle\sigma_x\sigma_y\rangle^0_\beta \;\equiv\; \inf_{B\subset\mathbb{Z}^d, |B|<\infty} \left\langle \left[\frac{1}{|B|}\sum_{x\in B}\sigma_x\right]^2 \right\rangle^0_\beta \tag{2.6}$$

which satisfies

$$\inf_{x\in\mathbb{Z}^d} \langle\sigma_0\sigma_x\rangle^0_\beta \;\leq\; \widetilde{M}_{LRO}(\beta) \;\leq\; M_{LRO}(\beta)\,. \tag{2.7}$$

It may be noted that whereas $m^*(\beta_c)$ provides direct information about the states at $\beta > \beta_c$, the monotonicity arguments of [81] imply that $M_{LRO}(\beta_c)$ provides direct information about the states at $\beta < \beta_c$ and, furthermore, the following relation holds.

Proposition 2.4. *For any translation invariant ferromagnetic Ising model on $\mathbb{Z}^d$: at all $\beta \geq 0$*

$$M_{LRO}(\beta) \;\leq\; m^*(\beta)^2 \tag{2.8}$$

with equality holding at values of β at which $P(\beta, 0)$ is continuously differentiable.

Our main general result is:

Theorem 2.5 ([8]). *For any ferromagnetic Ising model on $\mathbb{Z}^d$ whose coupling constants $(J_{x,y})_{x,y\in\mathbb{Z}^d}$ satisfy the conditions C1-C4: if*

$$\widetilde{M}_{LRO}(\beta) \;=\; 0 \tag{2.9}$$

then also

$$m^*(\beta_c) \;=\; 0\,, \tag{2.10}$$

and the system has only one Gibbs state at β_c.

The proof is based on the technique of the model's *random current* representation which was developed in [3] starting from the Griffiths-Hurst-Sherman switching lemma (which was earlier used in [62] for the GHS inequality). In this representation the onset of the Ising model's symmetry breaking is presented as a percolation transition in a system of random currents with constrained sources. The perspective that this picture offers has already shown itself to be of value in yielding a range of results for the model's critical behavior (c.f. [3, 4, 6, 101]). The incremental step taken here is to consider directly the limiting shift invariant infinite systems of random currents. This allows to add to the available tools arguments based on the 'uniqueness of infinite cluster' principle, which is of relevance to the question of continuity of the state at the model's critical temperature.

Continuity for Potts model. The Potts model is a model of random coloring of $\mathbb{Z}^2$ introduced as a generalization of the Ising model to more-than-two components spin systems. In this model, each vertex of $\mathbb{Z}^2$ receives a spin among q possible colors. The energy of a configuration is proportional to the number of frustrated edges, meaning edges whose endpoints have different spins. The model was introduced by Potts [91] (actually it was suggested to him by his adviser Domb). In two dimensions, it exhibits a reach panel of possible critical behaviors depending on the number of colors, and despite the fact that the model is exactly solvable (yet not rigorously for $q \neq 2$), the mathematical understanding of its phase transition remains restricted to a few cases (namely $q = 2$ and q large).

An extensive physics literature has been devoted to the question of continuity of the phase transition in the case of the Potts model. In the planar case, Baxter [17–19] used a mapping between the Potts model and solid-on-solid ice-models to compute the free energy at criticality and was able to predict that the phase transition was continuous for $q \leq 4$ and discontinuous for $q \geq 5$. While this computation gives a good insight on the behavior of the model, it relies on unproved assumptions which, forty years after their formulation, seem still very difficult to justify rigorously. In [46] we prove that the phase transition is continuous for $q \in \{2, 3, 4\}$ without any reference to unproved assumptions.

Most of our work is devoted to the study of the so-called random-cluster model. This model is a probability measure on edge configurations (each edge is declared open or closed) such that the probability of a configuration is proportional to

$$p^{\#\text{ open edges}}(1-p)^{\#\text{ closed edges}}q^{\#\text{ clusters}},$$

where clusters are maximal connected subgraphs, and $(p, q) \in [0, 1] \times \mathbb{R}_+$. For $q = 1$, the model is simply Bernoulli percolation.

Since its introduction by Fortuin and Kasteleyn [57], the random-cluster model has become an important tool in the study of phase transitions. The spin correlations of Potts models are rephrased as cluster connectivity properties of their random-cluster representations via the Edwards and Sokal coupling [52]. As a byproduct, properties of the random-cluster model can be transferred to the Potts model, and vice-versa.

While the understanding of critical Bernoulli percolation in 2 dimensions is now fairly well understood, the case of the random-cluster model remains mysterious. The long range dependency makes the model challenging to study probabilistically, and some of its most basic properties were not proved until recently. In [46] we derive several properties of the critical model, including a suitable generalization of the celebrated *Russo-Seymour-Welsh* theory available for Bernoulli percolation. This powerful tool enables us to prove several new results on the critical phase.

Our approach fits to the more general context of the study of conformally invariant planar lattice models. In the early eighties, physicists Belavin, Polyakov and Zamolodchikov postulated conformal invariance of critical planar statistical models [22, 23]. This prediction enabled physicists to harness Conformal Field Theory in order to formulate many conjectures on these models. From a mathematical perspective, proving rigorously the conformal invariance of a model (and properties following from it) constitutes a formidable challenge.

In recent years, the connection between discrete holomorphicity and planar statistical physics led to spectacular progress in this direction. Kenyon [70, 71], Smirnov [107] and Chelkak and Smirnov [35] exhibited discrete holomorphic observables in the dimer and Ising models and proved their convergence to conformal maps in the scaling limit. These results paved the way to the rigorous proof of conformal invariance for these two models. Other discrete observables have been proposed for a number of critical models, including self-avoiding walks and Potts models. While these observables are *not exactly discrete holomorphic*, their discrete contour integrals vanish, a property shared by discrete holomorphic functions.

It is a priori unclear whether this property is of any relevance for the models. Nevertheless, in the case of the self-avoiding walk, it was proved to be sufficient to compute the connective constant of the hexagonal lattice [49]. In our case, we also use parafermionic observables introduced independently in [60, 95, 107].

Definition of the models and main statements. Consider an integer $q \geq 2$ and a subgraph $G = (V_G, E_G)$ of the square lattice $\mathbb{Z}^2$. Here and below, V_G is the set of *vertices* of G and $E_G \subset V_G^2$ is the set of *edges*. For simplicity, the square lattice will be identified with its set of vertices, namely $\mathbb{Z}^2$. For two vertices $x, y \in V_G$, $x \sim y$ denotes the fact that $(x, y) \in E_G$.

Let $\tau \in \{1, \dots, q\}^{\mathbb{Z}^2}$. The *q-state Potts model on G with boundary conditions* τ is defined as follows. The space of configurations is $\Omega = \{1, \dots, q\}^{\mathbb{Z}^2}$. For a configuration $\sigma = (\sigma_x : x \in \mathbb{Z}^2) \in \Omega$, the quantity σ_x is called the *spin* at x (it is sometimes interpreted as being a color). The *energy* of a configuration $\sigma \in \Omega$ is given by the Hamiltonian

$$H_G^\tau(\sigma) := \begin{cases} -\sum_{\substack{x \sim y \\ \{x,y\} \cap G \neq \emptyset}} \delta_{\sigma_x, \sigma_y} & \text{if } \sigma_x = \tau_x \text{ for } x \notin V_G, \\ \infty & \text{otherwise.} \end{cases}$$

Above, $\delta_{a,b}$ denotes the Kronecker symbol equal to 1 if $a = b$ and 0 otherwise. The spin-

configuration is sampled proportionally to its Boltzmann weight: at an inverse-temperature β, the probability $\mu^{\tau}_{G,\beta}$ of a configuration σ is defined by

$$\mu^{\tau}_{G,\beta}[\sigma] \;:=\; \frac{\mathrm{e}^{-\beta H^{\tau}_G(\sigma)}}{Z^{\tau}_{G,\beta}} \quad \text{where} \quad Z^{\tau}_{G,\beta} \;:=\; \sum_{\sigma\in\Omega} \mathrm{e}^{-\beta H^{\tau}_G(\sigma)}$$

is the so-called *partition function* defined in such a way that the sum of the weights over all possible configurations equals 1.

Infinite-volume Gibbs measures can be defined by taking limits, as G tends to $\mathbb{Z}^2$, of finite-volume measures $\mu^{\tau}_{G,\beta}$. In particular, if $(i) := \tau$ denotes the constant configuration equal to $i \in \{1,\dots,q\}$, the sequence of measures $\mu^{(i)}_{G,\beta}$ converges, as G tends to infinity, to a Gibbs measure denoted by $\mu^{(i)}_{\mathbb{Z}^2,\beta}$. This measure is called the *infinite-volume Gibbs measure with monochromatic boundary conditions* i.

The Potts models undergoes a phase transition in infinite volume at a certain *critical inverse-temperature* $\beta_c(q) \in (0,\infty)$ in the following sense

$$\mu^{(i)}_{\mathbb{Z}^2,\beta}[\sigma_0 = i] = \begin{cases} \frac{1}{q} & \text{if } \beta < \beta_c(q), \\ \frac{1}{q} + m_\beta > \frac{1}{q} & \text{if } \beta > \beta_c(q). \end{cases}$$

The value $\beta_c(q)$ is computed in [20] and is equal to $\log(1+\sqrt{q})$ for any integer q (this value was previously known for $q = 2$ [88] and $q \geq 26$ [79]).

The phase transition is said to be *continuous* if $\mu^{(i)}_{\mathbb{Z}^2,\beta_c(q)}[\sigma_0 = i] = \frac{1}{q}$ and *discontinuous* otherwise. The main result is the following.

Theorem 2.6 ([46] Continuity of the phase transition for 2, 3 or 4 colors). *Let $q \in \{2,3,4\}$. Then for any $i \in \{1,\dots,q\}$, we have*

$$\mu^{(i)}_{\mathbb{Z}^2,\beta_c(q)}[\sigma_0 = i] = \tfrac{1}{q}.$$

This result was known in the $q = 2$ case. For two colors, the model is simply the Ising model. Onsager computed the free energy in [88] and Yang obtained a formula for the magnetization in [114]. In particular, this formula implies that the magnetization is zero at criticality. This results has been reproved in a number of papers since then. Let us mention a recent proof [113] not harnessing any exact integrability.

For q equal to 3 or 4, the result is new. Exact (yet non rigorous) computations performed by Baxter strongly suggest that the phase transition is continuous for $q \leq 4$, and discontinuous for $q > 4$. This result therefore tackles the whole range of q for which the phase transition is continuous. Let us mention that the technology developed in [46] is not restricted to the study of Potts models for $q \leq 4$: a property of Potts models with $q \geq 5$ colors witnessing ordering at criticality is also derived. Unfortunately, we were unable to show rigorously that the phase transition is discontinuous in the sense defined above.

In dimension $d \geq 3$, the phase transition is expected to be continuous if and only if $q = 2$. The best results in this direction are the following. On the one hand, the fact that the phase transition is continuous for the Ising model ($q = 2$) is known for any $d \geq 3$ [8] (in fact, the critical exponents are known to be taking their mean-field value for $d \geq 4$ [6]). On the other hand, mean-field considerations together with Reflection Positivity enabled [30] to prove that for any $q \geq 3$, the q-state Potts model undergoes a discontinuous phase transition above some dimension $d_c(q)$. Finally, Reflection Positivity can be harnessed to prove that for any $d \geq 2$, the phase transition is discontinuous provided q is large enough [74].

The random-cluster model. The proof of Theorem 2.6 is based on the study of a graphical representation of the Potts model, called the *random-cluster model*. Consider $p \in [0,1]$, $q > 0$ and a subgraph $G = (V_G, E_G)$ of the square lattice. A configuration ω is an element of $\Omega' = \{0,1\}^{E_G}$. An edge e with $\omega(e) = 1$ is said to be *open*, while an edge with $\omega(e) = 0$ is said to be *closed*. Two vertices x and y in V_G are said to be *connected* (this event is denoted by $x \longleftrightarrow y$) if there exists a sequence of vertices $x = v_1, v_2, \ldots, v_{r-1}, v_r = y$ such that (v_i, v_{i+1}) is an open edge for every $i < r$. A *connected component* of ω is a maximal connected subgraph of ω. Let $o(\omega)$ and $c(\omega)$ be respectively the number of open and closed edges in ω.

The *random-cluster measure on E_G with edge-weight p, cluster-weight q, and* free *boundary conditions* is defined by the formula

$$\phi^0_{G,p,q}\omega = \frac{p^{o(\omega)}(1-p)^{c(\omega)}q^{k_0(\omega)}}{Z^0_{G,p,q}},$$

where $k_0(\omega)$ is the number of connected components of the graph ω, and $Z^0_{G,p,q}$ is defined in such a way that the sum of the weights over all possible configurations equals 1. We also define the *random-cluster measure on E_G with edge-weight p, cluster-weight q, and* wired *boundary conditions* by the formula

$$\phi^1_{G,p,q}\omega = \frac{p^{o(\omega)}(1-p)^{c(\omega)}q^{k_1(\omega)}}{Z^1_{G,p,q}},$$

where $k_1(\omega)$ is the number of connected components of the graph ω, except that all the connected components of vertices in the vertex boundary ∂G, i.e. the set of vertices in V_G with less than four neighbors in G, are counted as being part of the same connected component. Again, $Z^1_{G,p,q}$ is defined in such a way that the sum of the weights over all possible configurations equals 1.

For $q \geq 1$, infinite-volume measures can be defined on $\mathbb{Z}^2$ by taking limits of finite-volume measures for graphs tending to $\mathbb{Z}^2$. In particular, the *infinite-volume random-cluster measure with free (resp. wired) boundary conditions* $\phi^0_{\mathbb{Z}^2,p,q}$ (resp. $\phi^1_{\mathbb{Z}^2,p,q}$) can be defined as the limit of the sequence of measures $\phi^0_{G,p,q}$ (resp. $\phi^1_{G,p,q}$) for $G \nearrow \mathbb{Z}^2$.

The random-cluster model with $q \geq 1$ undergoes a phase transition in infinite volume in the following sense. There exists $p_c(q) \in (0,1)$ such that

$$\phi^1_{\mathbb{Z}^2,p,q}[\omega]0 \longleftrightarrow \infty = \begin{cases} 0 & \text{if } p < p_c(q), \\ \theta^1(p,q) > 0 & \text{if } p > p_c(q), \end{cases}$$

where $\{0 \longleftrightarrow \infty\}$ denotes the event that 0 belongs to an infinite connected component. The value of $p_c(q)$ was recently proved to be equal to $\sqrt{q}/(1+\sqrt{q})$ for any $q \geq 1$ in [20]. The result was previously proved in [72] for Bernoulli percolation ($q = 1$), in [88] for $q = 2$ using the connection with the Ising model and in [80] for $q \geq 25.72$.

Similarly to the Potts model case, a notion of continuous/discontinuous phase transition can be defined: the phase transition is said to be *continuous* if

$$\phi^1_{\mathbb{Z}^2,p_c(q),q}[0 \longleftrightarrow \infty] = 0$$

and discontinuous otherwise. The following theorem is the *alter ego* of Theorem 2.6.

Theorem 2.7 ([46] Continuous phase transition for cluster-weight $1 \leq q \leq 4$). *Let* $q \in [1,4]$, *then* $\phi^1_{\mathbb{Z}^2,p_c(q),q}[0 \longleftrightarrow \infty] = 0$.

Let us now describe briefly the Edwards-Sokal coupling [52] and its application. Fix $q \geq 2$ integer. From a random-cluster configuration sampled according to $\phi^1_{G,p,q}$, color each component (meaning all the vertices in it) with one color chosen uniformly in $\{1,\dots,q\}$, except for the connected component containing the vertices in ∂G which receive color i. The law of the random coloring thus obtained is $\mu^{(i)}_{G,\beta}$, where $\beta = -\log(1-p)$. This coupling between the random-cluster model with integer cluster-weight and the Potts models enables to deduce Theorem 2.6 from Theorem 2.7 immediately. Proving Theorem 2.7 requires a much better understanding of the critical phase than the one available until now. Except for the $q = 1$, $q = 2$ and $q \geq 25.72$ cases, very little was known on critical random-cluster models. The following theorem provides new insight on the possible critical behavior of these models.

For an integer n, let Λ_n denote the *box* $[-n,n]^2$ *of size* n. An open path is a path of adjacent open edges (we refer to the next section for a formal definition). Let $0 \longleftrightarrow \partial\Lambda_n$ be the event that there exists an open path from the origin to the boundary of Λ_n. For a rectangle $R = [a,b] \times [c,d]$, let $\mathcal{C}_h(R)$ be the event that there exists an open path in R from $\{a\} \times [c,d]$ to $\{b\} \times [c,d]$.

Theorem 2.8. *Let* $q \geq 1$. *The following assertions are equivalent :*

P1) *(Absence of infinite cluster at criticality)* $\phi^1_{\mathbb{Z}^2,p_c,q}[0 \longleftrightarrow \infty] = 0$.

P2) $\phi^0_{[\mathbb{Z}^2,p_c,q} = \phi^1_{\mathbb{Z}^2,p_c,q}$.

P3) *(Infinite susceptibility)* $\chi^0(p_c,q) := \sum_{x\in\mathbb{Z}^2} \phi^0_{\mathbb{Z}^2,p_c,q}[0 \longleftrightarrow x] = \infty$.

P4) *(Sub-exponential decay for free boundary conditions)*

$$\lim_{n\to\infty} \tfrac{1}{n} \log \phi^0_{\mathbb{Z}^2,p_c,q}[0 \longleftrightarrow \partial\Lambda_n] = 0.$$

P5) *(RSW) For any* $\alpha > 0$, *there exists* $c = c(\alpha) > 0$ *such that for all* $n \geq 1$,

$$\phi^0_{[-n,(\alpha+1)n]\times[-n,2n],p_c,q}[\mathcal{C}_h([0,\alpha n] \times [0,n])] \geq c.$$

The previous result was previously known in a few cases:

- *Bernoulli percolation (random-cluster model with* $q = 1$*).* In such case P2) is obviously satisfied. Furthermore, Russo [100] proved that P1), P3) and P4) are all true (and therefore equivalent). Finally, P5) was proved by Russo [100] and Seymour-Welsh [104].
- *Random-cluster model with* $q = 2$. This model is directly related to the Ising via the Edwards-Sokal coupling.
- *Random-cluster model with* $q \geq 25.72$. In this case, none of the above properties are satisfied, as proved by using the Pirogov-Sinai technology [80].

3. Self-destructive percolation

Self-destructive percolation was introduced in [26] in an attempt to understand the behavior (and existence) of infinite-parameter forest fire model studied in physics literature. It may be formulated for both bond and site percolation, we choose to consider the latter. Fix some infinite connected graph G.

For $\delta, p \geq 0$ consider a regular site percolation configuration with intensity p. Close all sites contained in the possibly many infinite clusters; we say infinite clusters are "burned". Finally open every site with probability δ, independently of all previous choices. Call $\mathbb{P}_{p,\delta}$ the measure governing the configuration thus obtained and $\theta(p,\delta)$ the $\mathbb{P}_{p,\delta}$-probability that a given site (called the origin) is in an infinite cluster.

More formally, let $\mathbb{Z}^2$ denote the square lattice with vertices $V(\mathbb{Z}^2)$ (also called *sites*) and edges $E(\mathbb{Z}^2)$. For sites $x, y \in V(\mathbb{Z}^2)$ we write $x \sim y$, alternatively $(x,y) \in E(\mathbb{Z}^2)$, when $\|x-y\|_2 = 1$. Set $\Omega = \{0,1\}^{V(\mathbb{Z}^2)}$. We call an element $\omega \in \Omega$ a *configuration* and write $\{\omega(x) : x \in V(\mathbb{Z}^2)\}$ for its coordinates. A site x with $\omega(x) = 1$ is called *open* (or ω-open when the configuration needs to be specified), while one with $\omega(x) = 0$ is called *closed*.

For a configuration ω and $x, y \in V(\mathbb{Z}^2)$, we say x is *connected* to y in ω, and write $x \leftrightarrow^{\omega} y$, if there exists an ω-open path with endpoints x and y. We write $x \leftrightarrow^{\omega} \infty$ and say that x is *connected to infinity* if there exists an infinite ω-open path starting at x. Finally we write $x \leftrightarrow^{\omega} y$ and $x \leftrightarrow^{\omega} \infty$ for the negations of the above events. A *cluster* is a connected component of the graph induced by the open sites of $\mathbb{Z}^2$.

For $p \in [0,1]$, let $\mathbb{P}_p$ be the site percolation measure on $\mathbb{Z}^2$ with intensity p. That is $\mathbb{P}_p$ is the product measure on Ω with $\mathbb{P}_p(\omega(x) = 1) = p$ for all $x \in V(\mathbb{Z}^2)$. Finally let $p_c = \sup\{p \geq 0 : \mathbb{P}_p(0 \leftrightarrow^{\omega} \infty) = 0\}$. For $p > p_c$ it is well known that there exists $\mathbb{P}_p$-a.s. a unique infinite cluster.

Let $p \in [0,1]$ and consider a configuration ω chosen according to $\mathbb{P}_p$. We define a modification of ω, called $\bar{\omega}$, as follows.

For $x \in V(\mathbb{Z}^2)$,

$$\bar{\omega}(x) = \begin{cases} 1 & \text{if } \omega(x) = 1 \text{ and } x \leftrightarrow^{\omega} \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Let $\delta \geq 0$ and σ be a configuration chosen according to $\mathbb{P}_\delta$, independently of ω. The enhancement of $\bar{\omega}$ with intensity δ is $\bar{\omega}^\delta(x) = \bar{\omega}(x) \vee \sigma(x)$.

Let $\mathbb{P}_{p,\delta}$ denote the probability measure governing ω, σ and thus $\bar{\omega}$ and $\bar{\omega}^\delta$. To avoid confusion, when working with $\mathbb{P}_{p,\delta}$, we will usually state which configuration we refer to. When writing simply $\mathbb{P}_{p,\delta}(A)$ we mean $\mathbb{P}_{p,\delta}(\bar{\omega}^\delta \in A)$. Let

$$\theta(p,\delta) = \mathbb{P}_{p,\delta}(0 \leftrightarrow^{\bar{\omega}^\delta} \infty). \tag{3.2}$$

Note that $\mathbb{P}_{p,\delta}$ is increasing in δ, hence so is θ.

Let $\delta_c(p) = \sup\{\delta : \theta(p,\delta) > 0\}$ and let $p_c = p_c(G)$ denote the critical point for regular site percolation. Then it is easy to see that $\delta_c(p) = \frac{p_c - p}{1-p}$ for $p \leq p_c$. Hence self-destructive percolation is only interesting for $p > p_c$. In [26], it was conjectured that, for planar lattices, δ_c is uniformly bounded away from 0 when $p > p_c$.

The conjecture is somewhat surprising. When p is very close to p_c the infinite percolation cluster is very thin, and even after burning it, one may expect that opening only few sites suffices to obtain a new infinite cluster.

Recently in [2] was proven that, for non-amenable graphs G, the conclusion of the conjecture is false, i.e. that $\delta_c(p) \to 0$ as $p \to p_c$. The same has been shown in [1] for high dimensional lattices (more precisely for bond percolation on $\mathbb{Z}^d$ with d large enough).

In two dimensions it has been proved in [26, Prop. 3.1] that $\delta_c(p) > 0$ for any given $p > p_c$. This later was strengthened by van den Berg and de Lima [29] to the linear lower bound $\delta_c(p) \geq (p - p_c)/p$, but a bound which is non-zero and uniform in p could not be obtained. Finally, in [73] the afore-mentioned conjecture was proved:

Theorem 3.1 ([73]). *If G is planar and invariant under translation (by some $u \in \mathbb{R}^2 \setminus \{0\}$), rotation (of an angle $\varphi \in (0, \pi)$) and reflection with respect to a line, then there exists $\delta > 0$ such that for all $p > p_c(G)$, $\theta(p, \delta) = 0$.*

We mention that Theorem 3.1 also holds in the same form for bond percolation. In the present paper we will prove the theorem in the setting of site percolation on $\mathbb{Z}^2$. We will point out throughout the paper how to adapt our proof to other planar lattices.

Let us turn to the implications of Theorem 3.1. Let δ_c be the limit of $\delta_c(p)$ as $p \searrow p_c$. Theorem 3.1 together with the results in [28] shows that the function $(p, \delta) \to \theta(p, \delta)$ is continuous on the set $[0, 1]^2 \setminus \{p_c\} \times (0, \delta_c]$, while it is discontinuous on $\{p_c\} \times (0, \delta_c]$.

This result has important consequences for forest fires. Intuitively, an infinite-parameter forest fire is a process indexed by $t \geq 0$ defined as follows. At the initial time $t = 0$, all sites are closed. As t increases, sites open independently, at times distributed exponentially with rate 1. When an infinite cluster appears it is immediately burned (i.e. all its sites are closed). Then sites become open again at rate 1. It is not clear whether such a model actually exists. In [73] we also show that our results combined with those in [26] imply that infinite-parameter forest fires cannot be defined on two-dimensional lattices.

To avoid the problems of definition, one can investigate the N-parameter forest fire models with $N < \infty$. That is, we modify the dynamics above by burning clusters as soon as their 'size' reaches N. Our results with those of [27] provide some insight to the behavior of these processes. We find a behavior which is quite different compared to that of a mean field version of the forest fire model cf. [93].

4. Non-equilibrium phase transitions

Modern Statistical Mechanics offers a large and important class of driven-dissipative lattice systems that naturally evolve to a critical state, which is characterized by power-law distributions of the sizes of relaxation events (a paradigm example is the emergence of avalanches caused by small perturbations in sandpile models). In many mathematically interesting and physically relevant cases such systems are attracted to a stationary critical state without being specifically tuned to a critical point. In particular, it is believed that this phenomenon lies behind random fluctuations at the macroscopic scale, and creation of self-similar shapes in a variety of growth systems.

Due to strong non-locality of correlations and dynamic long-range effects, classical analytic and probabilistic techniques fail in most cases of interest, making the rigorous analysis of such systems a major mathematical challenge.

Studies of the above phenomenon are confined to very few models, and its conceptual understanding is extremely fragmented. Among theories which attempt to explain long-ranged spatio-temporal correlations, the physical paradigm called 'self-organized criticality' takes

its particular place [42, 67]. These are systems whose natural dynamics drives them towards, and then maintains them at the *edge* of stability [42]. However, from point of view of non-equilibrium statistical mechanics it is becoming increasingly evident that 'self-organized criticality' is related to conventional critical behavior, namely that of an *absorbing-state phase transition*. The known examples are variations of underlying non-equilibrium systems which actually do have a parameter and exhibit critical phenomena. The phase transition in these systems arises from a conflict between a spread of activity and a tendency for this activity to die out [43, 83], and the transition point separates an active and an absorbing phase in which the dynamics gets eventually extinct in any finite region.

Here we focus on two chief examples of conservative, infinite-volume systems which belong to the above mentioned family: the activated random walk model for reaction-diffusion and the stochastic sandpile model. We briefly describe the models.

The reaction-diffusion model is given by the following conservative particle dynamics in $\mathbb{Z}^d$. Each particle can be in one of two states: an active A-state, and a passive S-state. Each particle in the A-state performs a continuous-time random walk with jump rate $D_A = 1$. The jumps have a probability density $p(\cdot)$ on $\mathbb{Z}^d$ such that the set $\{z \in \mathbb{Z}^d : p(z) > 0\}$ generates the whole group $(\mathbb{Z}^d, +)$. Independently of anything else, each particle in the A-state turns to the S-state at a halting rate $\lambda > 0$. Once a particle is in the S-state, it stops moving, i.e., its jump rate is $D_S = 0$, and it remains in the S-state until the instant when another particle is present at the same vertex. At such an instant the particle which is in S-state flips to the A-state, giving the transition $A + S \to 2A$. A particle in the S-state stands still forever if no other particle ever visits the vertex where it is located. The catalyzed transition $A + S \to 2A$ and the spontaneous transition $A \to S$ represent the spread of activity versus a tendency for this activity to die out. According to these rules, the transition $A \to S$ *effectively* occurs if and only if, at the instant of such a transition, the particle *does not share* its vertex with another particle (the innocuous instantaneous transition $2A \to A+S \to 2A$ is not observed). Particles in the A-state do not interact among themselves. This system will be referred to as the model of Activated Random Walks (ARW).

In sandpile models the state of the system is represented by the number of particles $\eta(x) = 0, 1, 2, \dots$ at each vertex $x \in \mathbb{Z}^d$. The vertex x is *stable* when $\eta(x) < N_c$, for some threshold value N_c, and *unstable* when $\eta(x) \geqslant N_c$. Relaxation (update of the state) happens by toppling each unstable vertex, i.e., sending particles to its neighbors following a certain (deterministic or stochastic) rule. Here we study the following sandpile dynamics, which is a variation of Manna's model frequently considered in the physics literature [41, 84]. The threshold for stability of vertices is $N_c = 2$, and each unstable vertex topples after an exponentially-distributed time, sending 2 particles to neighbors chosen independently at random. We will refer to this model as the Stochastic Sandpile Model (SSM).[1]

For these systems, the relation between self-organized and ordinary criticality is understood as follows. On the one hand, 'self-organized criticality' appears in their parameter-free, finite-volume variation: particles are added to the bulk of a finite box, and absorbed at its boundary during relaxation. The particle addition happens at a *slow* rate, or *only after* the system globally stabilizes. In this dynamics, when the average density μ inside the box is too small, mass tends to accumulate. When it is too large, there is intense activity and a substantial number of particles is absorbed at the boundary. With this carefully designed mechanism, the model is attracted to a *critical state* with an average density given by

[1] In the so-called *Abelian Sandpile*, $N_c = 2d$, and an unstable vertex *deterministically* sends one particle to each of its $2d$ neighbors when toppling.

$0 < \mu_c < \infty$, though it was not explicitly tuned to this critical value. On the other hand, the corresponding conservative systems in infinite volume exhibit ordinary criticality in the sense that their dynamics fixate for $\mu < \mu_c$ and do not fixate for $\mu > \mu_c$, and moreover the *critical exponents* of the finite-volume addition-relaxation dynamics are related to those of the conservative dynamics in infinite volume.

The *critical behavior* of stochastic sandpiles seems to belong to the same *universality class* as the depinning of a linear elastic interface subject to random pinning potentials, roughly depicted by a nailed carpet being detached from the floor by an external force of critical intensity, where the rupture of each nail induces other ruptures nearby, giving rise to "avalanches".

The deterministic sandpile defines a universality class *sui generis*, and is marked by strong non-ergodic effects [43].

This seems to be due to the failure of the toppling procedure in eliminating certain microscopic symmetries in the configuration by the time the system becomes unstable, as a consequence of the existence of many toppling invariants.

In the stochastic sandpile models, by contrast, the addition-relaxation operators are themselves random, leading to a set of coupled polynomial equations [42, 102]. Their explicit solutions are not known in a general form, and very little can be said rigorously about the phase transition of such systems. The reaction-diffusion dynamics, in turn, might be even more apt to dispel microscopic symmetries, for not only the moves are random, but also the particles jump individually rather than in pairs.

The main pursuit in this framework is to describe the critical behavior, the scaling relations and critical exponents, and whether the critical density is the same as the long-time limit attained in the driven-dissipative version. These questions are however far beyond the reach of current techniques.

Despite of multiple efforts even the existence of phase transition for these two models was not know till recently.

Definition 4.1. We say that the system *locally fixates* if $\eta_t(x)$ - the number of particles at vertex x, is eventually constant for each x, otherwise we say that the system *stays active*.

Theorem 4.2 ([97]). *Consider the Stochastic Sandpile Model in the one-dimensional lattice $\mathbb{Z}$, with initial distribution ν given by i.i.d. Poisson random variables with parameter μ. There exists $\mu_c \in \left[\frac{1}{4}, 1\right]$ such that the system locally fixates a.s. if $\mu < \mu_c$, and stays active a.s. if $\mu > \mu_c$.*

Theorem 4.3 ([97]). *Consider the Activated Random Walk Model with nearest-neighbor jumps in the one-dimensional lattice $\mathbb{Z}$ with halting rate λ and with initial distribution ν given by i.i.d. Poisson random variables in $\mathbb{N}_0$ with parameter μ. There exists $\mu_c \in \left[\frac{\lambda}{1+\lambda}, 1\right]$ such that the system locally fixates a.s. if $\mu < \mu_c$ and stays active a.s. if $\mu > \mu_c$.*

Theorem 4.3 should be contrasted with the particular case of totally asymmetric jumps, for which it is known that $\mu_c = \frac{\lambda}{1+\lambda}$, so the lower bound is sharp. Note also that $\mu_c = 1$ when $\lambda = +\infty$. Theorems 4.2 and 4.3 remain true under more general hypotheses.

Remark 4.4. The methods and ideas used in [97] potentially could be extended to higher dimensions, however they come short due to technical difficulties controlling geometry of configuration. By using multi scale analysis techniques results analogous to the Theorems 4.2 and 4.3 were recently obtained in [106] for any dimension, provided the density of particles is sufficiently small.

Critical flow in one dimension. Here we consider the *flow process*, i.e., the process which counts the amount of particles which have passed through the origin 0. We find the scaling limit of this process for the biased particle-hole model in $\mathbb{Z}$ (see [32] for definition), which is given by the running maximum of a Brownian motion.

The biased particle-hole model is very close to ARW and a similar scaling limit should hold for the ARW with asymmetric walks at $\lambda = \infty$. It would be interesting to understand the scaling limit of totally-asymmetric walks with finite λ at critical density $\mu_c = \frac{\lambda}{1+\lambda}$, but we have not been able to find the correct description. The case of asymmetric walks and finite λ is much less clear, let alone that of symmetric walks.

Consider the particle-hole model with jump probabilities $p > \frac{1}{2}$ to the right and $q = 1-p$ to the left, and initial condition having mean $\mu = 1$ and positive finite variance σ^2. We define the flow process as

$$C_t := \text{number of particles which have passed through } 0 \text{ before time } t, \quad t \geq 0.$$

Let $(B_t)_{t\geq 0}$ be a one-dimensional Brownian motion started at 0 and $\tilde{B}_t := \max\{B_s : s \leq t\}$ denote its running maximum. The theorem below states that the scaling limit of the flow process $(C_t)_{t\geq 0}$ is $(\tilde{B}_t)_{t\geq 0}$. The *plateaux* of $\tilde{B}$ (given by excursions of B below its running maximum) correspond to the ever longer intervals of inactivity at the origin in the model. Moreover, the scale invariance $\tilde{B}_{L^2 t} \stackrel{d}{=} L\,\tilde{B}_t$ indicates that the amount of particles which pass through the origin before time t is of order $\sqrt{t}$, providing a critical exponent. The above observations are in agreement with the predictions of vanishing activity and non-fixation.

Theorem 4.5 ([32]). *For $d = 1$, let $v = p - q > 0$ denote the average speed of a moving particle in the particle-hole model. Assume $\mathbb{E}[\eta_0(0)] = 1$ and $\mathbb{E}[\eta_0(\boldsymbol{o})^2] = 1 + \sigma^2$ with $0 < \sigma < \infty$. Then the scaling limit of the flow process $(C_t)_{t\geq 0}$ is given by*

$$\left(\tfrac{1}{\sigma L}\, C_{\frac{L^2 t}{v}}\right)_{t\geq 0} \xrightarrow{d} \left(\tilde{B}_t\right)_{t\geq 0},$$

where $\xrightarrow{d}$ denotes convergence in distribution in $D[0,\infty)$ with the M_1-topology.

For many open problems we refer to [32, 44].

5. Dynamic phase transition and slow bond problem

In this section we address to the following question: how can localized defect, especially if it is small, affect the macroscopic behavior of a growth system? This is one of the fundamental questions in non-equilibrium growth: is the asymptotic shape changed (faceted) in the macroscopic neighborhood of such a defect at any value of its strength, or, when the defect is too weak, do the fluctuations of the bulk evolution become predominant and destroy the effects of the obstruction in such a way that its presence becomes macroscopically undetectable?

Such a vanishing presence of the macroscopic effect as a function of the strength of obstruction represents what is called *dynamic phase transition*. The existence of such a transition, its scaling properties, the shape of the density profile near the obstruction are among the most important issues.

Consider classical Ulam's problem of maximal increasing subsequence casted in the language of continuum last passage percolation: Let Π be a Poisson point process of intensity 1 on $\mathbb{R}^2$. For points $u = (0,0)$ and $u' = (n,n)$ let L_n denote the maximum number of points which can be collected along an increasing path from u to u'. We call L_n the *length* of a maximal path from $(0,0)$ to (n,n). It is well known (see [82, 112], and [9] for an alternative proof), that

$$\lim_{n\to\infty} \frac{\mathbb{E}L_n}{n} = 2. \tag{5.1}$$

Now, for $\lambda > 0$, let Σ_λ be a one dimensional poisson process of intensity λ on the line $x = y$ independent of Π. Let Π_λ be the point process obtained by superimposing Π and Σ_λ, i.e., a realisation of Π_λ is obtained by taking two independent realisations of Π and Σ_λ, and superimposing the point configurations. Let L_n^λ denote the maximum number of points of Π_λ on an increasing path from $(0,0)$ to (n,n). It is easy to observe that taking λ sufficiently large changes the law of large numbers for L_n^λ from that of L_n, i.e., for λ sufficiently large

$$\lim_{n\to\infty} \frac{\mathbb{E}L_n^\lambda}{n} > 2. \tag{5.2}$$

An important problem is whether there is a non-trivial phase transition in λ, i.e., whether for any $\lambda > 0$ the law of large numbers for L_n^λ differs from that of L_n, or there exists $\lambda_c > 0$, such that the law of large number for L_n^λ is same as that of L_n for $\lambda < \lambda_c$. Our main result settles this question:

Theorem 5.1 ([16]). *For every $\lambda > 0$,*

$$\lim_{n\to\infty} \frac{\mathbb{E}L_n^\lambda}{n} > 2. \tag{5.3}$$

We also consider a discrete last passage percolation on $\mathbb{Z}_+^2$, defined by associating with each vertex $x \in \mathbb{Z}_+^2$ a random variable $\xi_x \sim \exp(1)$, and ξ_x are i.i.d. for all $x \in \mathbb{Z}_+^2$. Let $\pi = \{x_0 = (0,0), x_1, \ldots, x_n = (n,n)\}$ be an oriented path connecting $(0,0)$ to (n,n). Define

$$L_n^1 = \max_\pi \sum_{i=0}^{n} \xi_{x_i};$$

It is well known that

$$\lim_{n\to\infty} \frac{\mathbb{E}L_n^1}{n} = 4 \tag{5.4}$$

This description also corresponds to totally asymmetric exclusion process $X(t)$ in continuous time, where $X(t) = (\eta_k(t))_{k=-\infty}^{\infty} \in \{0,1\}^{\mathbb{Z}}$, a particle $(\eta_k = 1)$ jumps with exponential rate to the right one step provided there is no particle at $k+1(\eta_{k+1} = 0)$, and where the initial configuration is $\mathbb{I}_{(-\infty,0]}(k)$ ("step initial conditions"). This form of simple exclusion process was introduced and studied by Rost, [99]. Now let us modify the distribution of passage times, by taking

$$\xi_{(x,y)} \sim \begin{cases} \exp(1) & \text{if } x \neq y \\ \exp(1-\epsilon) & \text{if } x = y. \end{cases} \tag{5.5}$$

and ask the same question: does the law of large numbers change for any $\epsilon > 0$. In TASEP representation this change corresponds to local modification of the dynamics: particles are jumping across the edges of $\mathcal{E}(\mathbb{Z}) \setminus \langle 0, 1\rangle$ with intensity 1, and the edge $\langle 0, 1\rangle$ is crossed at intensity $1 - \epsilon$. This model was proposed in [68, 69] by Janowsky and Lebowitz in their attempt to understand nonequilibrium stationary states of macroscopic systems. Clearly the rate decrease will increase the particle density to the immediate left of this "blockage" bond and decrease the density to its immediate right, but what is not obvious is that this perturbation may have global effect, in addition to local effects, in particular change the current of the system. The question got to prominence due to long-standing controversy based on numerical analysis, weather $\epsilon_c > 0$ or LLN for L_n^ϵ changes for any value $\epsilon > 0$, and became known as "slow bond problem". The difficulty comes from the fact that differently from local perturbations of equilibrium systems, where the effect of such perturbation is local, effect of any local perturbation in non-equilibrium systems carrying fluxes of conserved quantities is felt at large scales.

Mean-field theory analysis predicted an infinitely long traffic jam for any $\epsilon > 0$ and only a logarithmic depletion density pro- file near a slow bond [68]. In retrospect, based on finite size scaling analysis of simulation data the value $\epsilon_c \sim 0.2$ was predicted in [64]. Rigorously this question was addressed also in [37, 103]. Our second results confirms predictions made in [68]:

Theorem 5.2 ([16]). *In discrete last passage percolation model for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \frac{\mathbb{E}L_n^\epsilon}{n} > 4. \tag{5.6}$$

The above mentioned models belong to $1 + 1$ dimensional Kardar - Parisi - Zhang universality class. Part of the success relies on the fact that models we mention here lend themselves to the explicit computation of various quantities, such as growth speed and interface fluctuations. The proofs depend on exact algebraic computations, and it is not clear how to apply them as soon as the dynamics are modified in a non-translation-invariant way, with an exception of symmetrized cases [13].

6. Coffman-Gilbert conjecture

In this final section we discuss another prominent model - a classical greedy single-server system on the unit-length circle $\mathbb{R}/\mathbb{Z}$. Customers arrive following a Poisson process with rate λ. Each arriving customer chooses a position on $\mathbb{R}/\mathbb{Z}$ uniformly at random and waits for service. If there are no customers in the system, the server stands still. Otherwise, the server chooses the nearest waiting customer and travels in that direction at speed $v > 0$, ignoring any new arrivals. Upon reaching the position of such customer, the server stays there until service completion, which takes a random time T that is independent of the past configurations and has expectation μ^{-1}.

The above system was introduced by Coffman and Gilbert in 1987 [36], and since then became a paradigm example of a routing mechanism that depends on the system state. This is the so-called *greedy server*, due to the simple strategy of targeting the nearest customer.

It was conjectured in [36] that when $\lambda < \mu$, the greedy server on the circle should be a stable system for any $v > 0$. Since then, a number of related models have been proposed

and studied. Stability was verified only in case of light-traffic assumptions, i.e., for λ and μ fixed and v large enough, or for case of the greedy server on a discrete ring $\mathbb{Z}/n\mathbb{Z}$.

However, these approximations were unable to identify and tackle the main difficulty of this system, which is is due to the interplay between the server's motion and the environment of waiting customers that surround it. This interplay is given by the interaction resulting from the choice of the next customer and the removal of those who have been served.

The main difficulty in studying greedy server systems in continuous spaces is due to the interplay between the server's motion and the environment of waiting customers that surround it. This interplay is given by the interaction resulting from the greedy choice of the next customer and the removal of those who have been served. The server's path is *self-repelling*, since the removal of already served customers makes it less likely for the greedy server to take the next step back into the recently visited regions.

In some well-known examples of self-repelling motions, the self-interaction comes from an explicit prescription of the distribution of next step in terms of the past occupation times. For the excited random walks [25], perturbed Brownian motions [33, 34, 39, 40, 90], and excited Brownian motions [92], whenever there is a drift, it is pushing the motion in a certain fixed direction. For the random walk avoiding its past convex hull [12, 115] and the prudent walk [21, 31], there is a growing forbidden region containing the previous trajectory, which strongly pushes the motion outwards.

For the greedy server, as well as for the class of other models, such as the true self-avoiding walk [109, 110], the true self-repelling motion [111], and the Brownian motion with repulsion [87], there is a mixture of information, and "self-repulsion" does not immediately imply "repulsion towards ∞", since the particle is allowed to cross its past path, receiving contradictory signals from different directions. In fact, some of the latter models are recurrent and some are transient.

It was clear since these models were introduced that they could not be treated via standard methods and tools. Still nowadays, a lot remains to be understood even in dimension $d = 1$, and, despite the existence of a few scattered techniques that have proved useful in particular situations, this rich field of study lacks a systematic basis.[2]

Few words on previous results. Stability was verified for the greedy server on $\mathbb{R}/\mathbb{Z}$ under light-traffic assumptions [77], and for the greedy server on a discrete ring $\mathbb{Z}/n\mathbb{Z}$ [58, 59, 85], see below. It was also shown for several related models, including a class of non-greedy policies [76], a gated-greedy variant on convex spaces [11], and random non-greedy servers on general spaces [10]. See [96] and references therein for a recent review.

The *light-traffic* regime is given by

$$\lambda \left(\frac{1}{\mu} + \frac{1}{2v} \right) < 1.$$

This regime was studied in [77], particularly the limit $\lambda \to 0$ for which the first terms of Taylor expansions of some performance measurements were computed. A simple coupling argument works for proving stability under light-traffic assumption.

On the other hand, stability under the general condition $\lambda < \mu$ is known to hold for the *polling server* on $\mathbb{R}/\mathbb{Z}$, i.e., the server whose strategy is to always travel in the same direction. In [75] this fact was proven using a decomposition of the set of waiting customers

[2] Except for the family of universality classes given by the Schramm-Loewner Evolutions, which include 2-dimensional loop-erased random walk and several other models.

into a collection of Galton-Watson trees that turn out to be subcritical for $\lambda < \mu$. This decomposition provides a detailed description of the busy cycles (sequence of configurations observed between two consecutive regeneration times) and the stationary state.

Simulations indicate that under heavy traffic conditions the greedy server dynamics resembles that of the polling server [36]. This suggests that a possible strategy for proving stability of the greedy server might be to adapt the above argument. In this case the first step would be to understand its local behavior, and a natural approach is to consider a system on an infinite line. A model on $\mathbb{Z}$ was studied in [78], where it is shown that the server is eventually going to move in a fixed random direction.

Yet, discrete models have not been able to grasp the microscopic nature of the greedy mechanism in continuous space, neither on $\mathbb{Z}$ nor on $\mathbb{Z}/n\mathbb{Z}$, and there are major obstacles in extrapolating any approach based on a discrete approximation. This difficulty is due to the *self-interaction* of the server's path at the *microscopic level*, which takes place because the server's trajectory influences the set of waiting customers and at the same time is determined by the latter.

Stochastic evolution of profiles. To address the issues mentioned above, we consider a representation of the customers environment which reflects its randomness as perceived by the server.

More precisely, we only want to learn the information that is necessary and sufficient to determine the next movement, and the positions of further waiting customers should remain unknown. Each time the server has to scan the system state to determine the position of the next target, we acquire exactly two pieces of information: the presence of a customer at that position and the absence of any other customer at smaller distances.

The arrivals are represented by a space-time Poisson Point Process $\nu \subseteq (\mathbb{R}/\mathbb{Z}) \times \mathbb{R}$, and in this approach one is ignoring the points of ν that have not yet influenced the server's trajectory. One can think of this scheme as re-sampling the set of waiting customers at each departure time, according to the appropriate conditional distribution. The latter is given in terms of the space-time region where the configuration ν has not been revealed. In this setting the state of the system is given by the positions of the server and the current customer, plus the profile corresponding to the boundary of this region where ν is unknown. The knowledge of this triplet determines the distribution of its future without the need of any further information from the past, yielding a Markovian evolution.

Definition 6.1. We say that t is a *regeneration time* if the system becomes empty at time t, i.e., if there is one customer at time $t-$ and no customers at time $t+$. Let $\tau_\emptyset := \inf\{t > 0 : t \text{ is a regeneration time}\}$. We say that the system is *recurrent* if, starting from the empty state $\emptyset$, there will be a.s. a regeneration time, i.e., $\mathbb{P}^\emptyset[\tau_\emptyset < \infty] = 1$. We say that the system is *stable*, or *positively recurrent*, if $\mathbb{E}^\emptyset[\tau_\emptyset] < \infty$.

Our main theorem settles Coffman-Gilbert conjecture in the dimension one:

Theorem 6.2 ([98]). *Suppose that the distribution of the service time T is geometric, exponential, or deterministic. For any $\lambda < \mu$ and any $v > 0$, the greedy server on the circle is stable.*

Heuristics of the proof is following: If the server is busy most of the time, the system must be stable, since in average the service time is smaller than the inter-arrival time. The fundamental problem in showing stability is therefore the possibility that the server spend a long time zigzagging on regions with low density of customers, due to a trapping configuration produced by the stochastic dynamics.

For the analogous model on the real line this cannot be the case: the server may zigzag for a finite period of time, but it is bound to eventually choose a direction and head that way [61].

On the same grounds, since the greedy routing mechanism is *local*, this can neither be the case on the circle – at least *until the server realizes that it is not operating on the infinite line*.

Suppose we are given a configuration where the circle is crowded of waiting customers, and, from this point on, our goal is to alleviate this situation. We would like to say that, with high probability, after a short time the server will choose a direction and then cope with its workload as the polling server would.

There are two situations where the server may feel that it is on the circle rather than on the line. First, if it arrives at a given point x for the second time after performing a whole turn on the circle, it will encounter an environment that has been affected by its previous visit. This is not a serious problem, because if it happens it will imply that all the customers which were initially present will have then been served, and typically the server will have served more customers than new ones will have arrived.

The second difference is what poses a real issue. The server has a tendency to go into regions that have been *least recently visited*, since in these regions the average interdistance between customers is smaller, and they have bigger chance to attract the server via its greedy mechanism. This is indeed how transience is proved on $\mathbb{R}$. Let us call the *age* of a point in space the measurement in time units of how recently it was visited by the server in the past. On the line, the age is minimal at the server's position, and *increases as we go further away from the server*. The new regions encountered thus become older and older, and the server surrenders to the fact that the cleared regions it is leaving behind cannot compete with the old regions ahead.

However, this is not true on the circle: the age profile cannot increase indefinitely. This gives rise to the possibility of the following tricky scenario. Imagine that on a tiny region around some point x the system is much older than on any close neighborhood. When the server enters this region, it will take a very long time to finish with all the waiting customers. After finishing with all these customers tightly packed in space, there will no longer be a strong difference between the ages ahead and behind the server, who may end up going back to the region that has just been cleared, invalidating the argument.

Open Problem. Under the assumption $\lambda < \mu$, show stability of the server on the two dimensional torus for any $v > 0$.

References

[1] D. Ahlberg, H. Duminil-Copin, G. Kozma, and V. Sidoravicius, *Seven-dimensional forest fires*, preprint, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, arXiv:1302.6872, 2013.

[2] D. Ahlberg, V. Sidoravicius, and J. Tykesson, *Bernoulli and self-destructive percolation on non-amenable graphs*, Electronic communications in Probability, preprint, arXiv:1302.6870, 2013.

[3] M. Aizenman, *Geometric analysis of ϕ^4 fields and Ising models*, Comm. Math. Phys. **86** (1982), no. 1, 1–48.

[4] M. Aizenman, D.J. Barsky, and R. Fernandez, *The phase transition in a general class of Ising-type models is sharp*, J. Stat. Phys. **47** (1987), no. 3-4, 343–374.

[5] M. Aizenman, J.T. Chayes, L. Chayes, and C.M. Newman, *Discontinuity of the magnetization in one-dimensional $1/|x-y|^2$ Ising and Potts models*, J. Stat. Phys. **50** (1988), no. 1-2, 1–40.

[6] M. Aizenman and R. Fernandez, *On the critical behavior of the magnetization in high-dimensional Ising models*, J. Stat. Phys. **44** (1986), no. 3-4, 393–454.

[7] M. Aizenman and C.M. Newman, *Uniqueness of the Infinite Cluster and Continuity of Connectivity Functions for Short and Long Range Percolation*, Commun. Math. Phys. **107** (1986), 505–531.

[8] M. Aizenman, H. Duminil-Copin, and V. Sidoravicius, *Random Currents and Continuity of Ising Model's Spontaneous Magnetization*, Comm. Math. Phys. Preprint arXiv:1311.1937, 2013.

[9] D. Aldous and P. Diaconis, *P. Hammersley's interacting particle process and longest increasing subsequences*, Probab. Theory Related Fields **103** (1995), no. 2, 199–213.

[10] E. Altman and S. Foss, *Polling on a space with general arrival and service time distribution*, Oper. Res. Lett. **20** (1997), 187–194.

[11] E. Altman and H. Levy, *Queueing in space*, Adv. in Appl. Probab. **26** (1994), 1095–1116.

[12] O. Angel, I. Benjamini, and B. Virág, *Random walks that avoid their past convex hull*, Electron. Comm. Probab. **8** (2003), pp. 6–16.

[13] J. Baik and E.M. Rains, *Symmetrized random permutations*, In MSRI Publications 40: Random Matrix Models and Their Applications, ed. P. Bleher and A. Its, Cambridge, 2001.

[14] D.J. Barsky, G.R. Grimmett, and C.M. Newman, *Dynamic renormalization and continuity of the percolation transition in orthants*, In Spatial stochastic processes, volume 19 of Progr. Probab., pages 37–55. Birkhäuser Boston, Boston, MA, 1991.

[15] ———, *Percolation in half-spaces: equality of critical densities and continuity of the percolation probability*, Probab. Theory Related Fields **90**(1) (1991), 111–148.

[16] R. Basu, V. Sidoravicius, A. Sly, *Last Passage Percolation with a Defect Line and Slow Bond Problem*, Preprint 2014.

[17] R.J. Baxter, *Generalized ferroelectric model on a square lattice*, Studies in Appl. Math. **50** (1971), 51–69.

[18] ———, *Potts model at the critical temperature*, Journal of Physics C: Solid State Physics, 6(23):L445, 1973.

[19] _____, *Exactly solved models in statistical mechanics*, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1989. Reprint of the 1982 original.

[20] V. Beffara and H. Duminil-Copin, *Critical point and duality in planar lattice models*, 2012.

[21] V. Beffara, S. Friedli, and Y. Velenik, *Scaling limit of the prudent walk*, Electron. Commun. Probab., 15 (2010), 44–58.

[22] A.A. Belavin, A.M. Polyakov, and A.B. Zamolodchikov. *Infinite conformal symmetry of critical fluctuations in two dimensions*, J. Statist. Phys. **34**(5-6) (1984), 763–774.

[23] _____, *Infinite conformal symmetry in two-dimensional quantum field theory*, Nuclear Phys. B, 241(2):333–380, 1984.

[24] I. Benjamini, R. Lyons, Y. Peres, and O. Schramm, *Critical percolation on any nonamenable group has no infinite clusters*, The Annals of Probability, 27(3):1347–1356, 1999.

[25] I. Benjamini and D.B. Wilson, *Excited random walk, Electron*, Comm. Probab., 8 (2003), pp. 86–92.

[26] J. van den Berg and R. Brouwer, *Self-destructive percolation*, Random Structures and Algorithms 24 (2004), no. 4, 480–501.

[27] _____, *Self-organized forest-fires near the critical time*, Comm. Math. Phys. 267 (2006), no. 1, 265–277.

[28] J. van den Berg, R. Brouwer, and B. Vagvolgyi, *Box-Crossings and Continuity Results for Self-Destructive Percolation in the Plane*, In and Out of Equilibrium 2, vol. 60, Birkhauser Basel, 2008, pp. 117–135.

[29] J. van den Berg and B.N.B. de Lima, *Linear lower bounds for c(p) for a class of 2D self-destructive percolation models*, Random Structures & Algorithms **34** (2009), no. 4, 520–526.

[30] M. Biskup, L. Chayes, and N. Crawford, *Mean field driven first-order phase transitions in systems with long- range interactions*, J. Stat. Phys. **122** (6) (2006), 1139–1193.

[31] M. Bousquet-Melou, *Families of prudent self-avoiding walks*, J. Combin. Theory Ser. A **117** (2010), 313–344.

[32] M. Cabezas, L. Rolla, V. Sidoravicius, *Non-equilibrium Phase Transitions: Activated Random Walks at Criticality*, Journal of Statistical Physics (2014), arXiv:1307.4450.

[33] P. Carmona, F. Petit, and M. Yor, *Beta variables as times spent in* $[0, \infty[$ *by certain perturbed Brownian motions*, J. London Math. Soc. (2), **58** (1998), pp. 239–256.

[34] L. Chaumont and R.A. Doney, *Pathwise uniqueness for perturbed versions of Brownian motion and reflected Brownian motion*, Probab. Theory Related Fields **113** (1999), 519–534.

[35] D. Chelkak and S. Smirnov, *Universality in the 2D Ising model and conformal invariance of fermionic observables*, Invent. Math. **189**(3) (2012), 515–580.

[36] J.E.G. Coffman and E. N. Gilbert, *Polling and greedy servers on a line*, Queueing Systems Theory Appl. **2** (1987), 15–145.

[37] P. Covert and F. Rezakhanlou, *Hydrodynamic Limit for Particle Systems with Nonconstant Speed Parameter*, Journal of Stat. Phys. **88**, No. 1/2, 1997.

[38] M. Damron, C.M. Newman, and V. Sidoravicius, *Absence of site percolation at criticality in* $\mathbb{Z}^2 \times \{0,1\}$, Random Structures and Algorithms, Preprint arXiv:1211.4138, 2012.

[39] B. Davis, *Weak limits of perturbed random walks and the equation* $Y_t = Bt + \alpha \sup\{Y_s : s \le t\} + \beta \inf\{Ys : s \le t\}$, Ann. Probab. **24** (1996), 2007–2023.

[40] ———, *Brownian motion and random walk perturbed at extrema*, Probab. Theory Related Fields **113** (1999), 501–518.

[41] D. Dhar, *The abelian sandpile and related models*, Physica A **263** (1999), 4–25.

[42] ———, *Theoretical studies of self-organized criticality*, Physica A 369, 29–70 (2006)

[43] R. Dickman, *Nonequilibrium phase transitions in epidemics and sandpiles*, Physica A **306** (2002), 90–97.

[44] R. Dickman, L.T. Rolla, and V. Sidoravicius, *Activated random walkers: facts, conjectures and challenges*, J. Stat. Phys. **138** (2010), 126–142.

[45] H. Duminil-Copin, V. Sidoravicius, and V. Tassion, *Absence of infinite cluster for critical Bernoulli percolation on slabs*, Preprint arXiv:1401.7130.

[46] ———, *Continuity of the phase transition for planar Potts models with* $1 \le q \le 4$, Preprint. 2014

[47] H. Duminil-Copin. *Parafermionic observables and their ap- plications to planar statistical physics models*, volume 25 of Ensaios Matematicos. Brazilian Mathematical Society, 2013.

[48] H. Duminil-Copin, V. Sidoravicius, and V. Tassion. *Continuous phase transition for planar Potts models with 1–q–4*, Preprint, 2014.

[49] H. Duminil-Copin and S. Smirnov. *The connective constant of the honeycomb lattice equals* $\sqrt{2+\sqrt{2}}$, Ann. of Math. (2) **175**(3) (2012), 1653–1665.

[50] R.L. Dobrushin, *Prescribing a system of random variables by the help of conditional distributions*, Prob. Theo. and its App. **15** (1970), 469–497.

[51] F.J. Dyson, *Existence of a phase-transition in a one-dimensional Ising ferromagnet*, Comm. Math. Phys. **12** (1969), no. 2, 91–107.

[52] R. G. Edwards and A. D. Sokal. *Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm*, Phys. Rev. D (3) **38** (6) (1988), 2009–2012.

[53] A. Fey, L. Levine, and D.B. Wilson, *Approach to criticality in sandpiles*, Phys. Rev. E **82**, 031121 (2010).

[54] _____, *Driving sandpiles to criticality and beyond*, Phys. Rev. Lett. **104**, 145703 (2010).

[55] A. Fey-den Boer and F. Redig, *Organized versus self-organized criticality in the Abelian sandpile model*, Markov Process. Relat. Fields **11** (2005), 425–442.

[56] M.E. Fisher, *Critical temperatures of anisotropic Ising lattices II*, general upper bounds, Phys. Rev. **162** (1967), 480.

[57] C.M. Fortuin and P.W. Kasteleyn. *On the random-cluster model. I. Introduction and relation to other models*, Physica **57** (1972), 536–564.

[58] S. Foss and G. Last, *Stability of polling systems with exhaustive service policies and state- dependent routing*, Ann. Appl. Probab. **6** (1996), 116–137.

[59] _____, *On the stability of greedy polling systems with general service policies*, Probab. Engrg. Inform. Sci. **12** (1998), 49–68.

[60] E. Fradkin and L. P Kadanoff, *Disorder variables and para-fermions in two-dimensional statistical mechanics*, Nuclear Physics B **170**(1) (1980), 1–15.

[61] S. Foss, L.T. Rolla, and V. Sidoravicius, *Transience of a server with greedy strategy on the real line*, Annals of Probability (to appear). arXiv:1111.4846v3, 2011.

[62] R.B. Griffiths, C.A. Hurst, and S. Sherman, *Concavity of mag- netization of an ising ferromagnet in a positive magnetic field*, J. Math. Phys. **11** (1970), 790.

[63] G.R. Grimmett and J.M. Marstrand, *The supercritical phase of percolation is well behaved*, Proceedings of the Royal Soci- ety of London. Series A: Mathematical and Physical Sciences, 430(1879):439, 1990.

[64] M. Ha, J. Timonen, and M. den Nijs, *Queuing transitions in the asymmetric simple exclusion process*, Phys. Rev. E 68 (2003).

[65] T. Hara and G. Slade, *Mean-field behaviour and the lace expansion*, NATO ASI Series C Math. and Physical Sciences-Advanced Study Institute, **420**(1994), 87–122.

[66] T.E. Harris, *A lower bound for the critical probability in a certain percolation process*, Math. Proceedings of the Cambridge Philosophical Society, **56**(01) (1960), 13–20.

[67] H. Hinrichsen, *Non-equilibrium critical phenomena and phase transitions into absorbing states*, Adv. Phys. **49** (2000), 815–958.

[68] S. Janowsky and J. Lebowitz, *Finite size effects and shock fluctuations in the asymmetric simple exclusion process*, Phys. Rev. A **45** (1992), 618-625.

[69] S. Janowsky and J. Lebowitz, *Exact results for the asymmetric simple exclusion process with a blockage*, J. Star. Phys. **77** (1994), 35-51.

[70] R. Kenyon, *Conformal invariance of domino tiling*, Ann. Probab. **28**(2) (2000), 759–795.

[71] ———, *Dominos and the Gaussian free field*, Ann. Probab. **29** (3) (2001), 1128–1137.

[72] H. Kesten, *The critical probability of bond percolation on the square lattice equals 1*, Comm. Math. Phys. **74**(1) (1980), 41–59.

[73] D. Kiss, I. Manolescu, and V. Sidoravicius, *Planar lattices do not recover from forest fires*, arXiv:1312.7004, 2013.

[74] R. Kotecky and S.B. Shlosman, *First-order phase transitions in large en- tropy lattice models*, Comm. Math. Phys. **83**(4) (1982), 493–515.

[75] D.P. Kroese and V. Schmidt, *A continuous polling system with general service times*, Ann. Appl. Probab. **2** (1992), 906–927.

[76] ———, *Single-server queues with spatially distributed arrivals*, Queueing Systems Theory Appl. **17** (1994), 317–345.

[77] ———, *Light-traffic analysis for queues with spatially distributed arrivals*, Math. Oper. Res. **21** (1996), 135–157.

[78] I.A. Kurkova and M.V. Menshikov, *Greedy algorithm, $\mathbb{Z}$ case*, Markov Process. Related Fields **3** (1997), 243–259.

[79] L. Laanait, A. Messager, and J. Ruiz. *Phases coexistence and surface ten- sions for the Potts model*, Comm. Math. Phys. **105**(4) (1986). 527–545.

[80] L. Laanait, A. Messager, S. Miracle-Sole, J. Ruiz, and S. Shlosman, *Interfaces in the Potts model. I. Pirogov-Sinai theory of the Fortuin-Kasteleyn representation*, Comm. Math. Phys. **140**(1) (1991), 81–91.

[81] J. Lebowitz, *Coexistence of phases in Ising ferromagnets*, J. Stat. Phys. **16** (1977), no. 6, 463–476.

[82] B.F. Logan and L.A. Shepp. *A variational problem for random Young tableaux*, Advances in Math. **26** (1977), 206–222.

[83] S. Lubeck, *Universal scaling behavior of non-equilibrium phase transitions*, Int. J. Mod. Phys. B **18** (2004), 3977–4118.

[84] S.S. Manna, *Two-state model of self-organized criticality*, J. Phys. A, Math. Gen. **24**, L363–L369 (1991).

[85] R. Meester and C. Quant, *Stability and weakly convergent approximations of queueing systems on a circle*, http://citeseerx.ist.psu.edu/viewdoc/summary, 1999.

[86] Meester, R.; Quant, C., *Connections between 'self-organised' and 'classical' criticality*, Markov Process. Relat. Fields **11** (2005), 355–370.

[87] T. Mountford and P. Tarrès, *An asymptotic result for Brownian polymers*, Ann. Inst. Henri Poincaré Probab. Stat. **44** (2008), 29–46.

[88] L. Onsager, *Crystal statistics. I. A two-dimensional model with an order-disorder transition*, Phys. Rev. (2) **65** (1944), 117–149.

[89] R. Peierls, *On Ising's model of ferromagnetism*, Math. Proc. Camb. Phil. Soc. **32** (1936), 477–481.

[90] M. Perman and W. Werner, *Perturbed Brownian motions*, Probab. Theory Related Fields **108** (1997), 357–383.

[91] R.B. Potts, *Some generalized order-disorder transformations*, In Proceedings of the Cambridge Philosophical Society **48**, pp 106–109. Cambridge Univ Press, 1952.

[92] O. Raimond and B. Schapira, *Excited Brownian motions*, ALEA Lat. Am. J. Probab. Math. Stat. **8** (2011), 19–41.

[93] B. Rath and B. Toth, *Erdos-Renyi randomgraphs+forestfires=self-organized criticality*, Electron. J. Probab. **14** (2009), no. 45, 1290–1327.

[94] Redig, F., *Mathematical aspects of the abelian sandpile model*, In: Bovier, A., Dunlop, F., van Enter, A., den Hollander, F., Dalibard, J. (eds.) Mathematical Statistical Physics' Session LXXXIII. Lecture Notes of the Les Houches Summer School, pp. 657–730. Else- vier, Amsterdam (2006).

[95] V. Riva and J. Cardy, *Holomorphic parafermions in the Potts model and stochastic Loewner evolution*, J. Stat. Mech. Theory Exp. (12):P12001, 19 pp. (electronic), 2006.

[96] L. Rojas-Nandayapa, S. Foss, and D. P. Kroese, *Stability and performance of greedy server systems: A review and open problems*, Queueing Syst. **68** (2011), pp. 221–227.

[97] Rolla, L.; Sidoravicius, V., *Absorbing-state phase transition for stochastic sandpiles and activated random walks*, Invent. Math. **188** (2012), no. 1, 127–150.

[98] L.T. Rolla and V. Sidoravicius, *Proof of Coffman-Gilbert Conjecture: Stability of the Greedy algorithm on the Circle*, arXiv:1112.2389, 2012

[99] H. Rost, *Non-Equilibrium Behaviour of a Many Particle Process: Density Profile and Local Equilibria*, Zeitschrift f. Warsch. Verw. Gebiete **58** (1981), 41–53.

[100] L. Russo, *A note on percolation*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **43**(1) (1978), 39–48.

[101] A. Sakai, *Lace expansion for the Ising model*, Comm. Math. Phys. **272** (2007), no. 2, 283–344.

[102] T. Sadhu and D. Dhar, *Steady state of stochastic sandpile models*, J. Stat. Phys. **134** (2009), 427–441.

[103] T. Seppalainen, *Hydrodynamic Profiles for the Totally Asymmetric Exclusion Process with a Slow Bond*, Journal of Statistical Physics, **102**, Nos. 1/2, 2001

[104] P.D. Seymour and D.J.A. Welsh, *Percolation probabilities on the square lattice*, Ann. Discrete Math., **3** (1978), 227–245. Advances in graph theory (Cambridge Combinatorial Conf., Trinity College, Cambridge, 1977).

[105] V. Sidoravicius, *Perplexing world of critical systems*, Preprint 2014.

[106] V. Sidoravicius and A. Teixeira, *Phase transition for activated random walks*, Preprint 2014.

[107] S. Smirnov, *Conformal invariance in random cluster models. I. Holomorphic fermions in the Ising model*, Ann. of Math. (2), **172**(2) (2010), 1435–1467.

[108] D.J. Thouless, *Long-range order in one-dimensional Ising systems*, Physical Review **187** (1969), 732–733.

[109] B. Toth, *The 'true' self-avoiding walk with bond repulsion on Z: limit theorems*, Ann. Probab. **23** (1995), pp. 1523–1556.

[110] _____, *Self-interacting random motions—a survey*, in Random walks (Budapest, 1998), vol. 9 of Bolyai Soc. Math. Stud., Janos Bolyai Math. Soc., Budapest, 1999, pp. 349–384.

[111] B. Toth and W. Werner, *The true self-repelling motion*, Probab. Theory Related Fields **111** (1998), pp. 375–452.

[112] A.M. Vershik and S.V. Kerov, *Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tables*, Soviet Math. Dokl. **18** (1977), 527–531. Translation of Dokl. Acad. Nauk. SSSR **233** (1977), 1024–1027.

[113] W. Werner, *Percolation et modèle d'Ising, volume 16 of Cours Spécialisés [Specialized Courses]*, Société Mathématique de France, Paris, 2009.

[114] C.N. Yang, *The spontaneous magnetization of a two-dimensional Ising model*, Phys. Rev. (2), **85** (1952), 808–816.

[115] M.P.W. Zerner, *On the speed of a planar random walk avoiding its past convex hull*, Ann. Inst. H. Poincare Probab. Statist. **41** (2005), 887–900.

IMPA, Estr. Dona Castorina 110, Jardim Botanico, Cep 22460-320, Rio de Janeiro, RJ, Brasil
E-mail: vladas@impa.br

Aggregation and minimax optimality in high-dimensional estimation

Alexandre B. Tsybakov

Abstract. Aggregation is a popular technique in statistics and machine learning. Given a collection of estimators, the problem of linear, convex or model selection type aggregation consists in constructing a new estimator, called the aggregate, which is nearly as good as the best among them (or nearly as good as their best linear or convex combination), with respect to a given risk criterion. When the underlying model is sparse, which means that it is well approximated by a linear combination of a small number of functions in the dictionary, the aggregation techniques turn out to be very useful in taking advantage of sparsity. On the other hand, aggregation is a general way of constructing adaptive nonparametric estimators, which is more powerful than the classical methods since it allows one to combine estimators of different nature. Aggregates are usually constructed by mixing the initial estimators or functions of the dictionary with data-dependent weights that can be defined is several possible ways. An important example is given by aggregates with exponential weights. They satisfy sharp oracle inequalities that allow one to treat in a unified way three different problems: Adaptive nonparametric estimation, aggregation and sparse estimation.

Mathematics Subject Classification (2010). Primary 62G05; Secondary 62J07.

Keywords. High-dimensional model, aggregation, sparsity, oracle inequality, minimax estimation, exponential weights.

1. Introduction

Aggregation of estimators in the regression model has been studied starting from [7, 27, 33, 39]. In this paper, we focus on the connection between aggregation and high-dimensional statistics. In particular, we show that some aggregation techniques, such as exponential weighting, achieve minimax rates in high-dimensional problems with sparsity constraints in an adaptive way. The results obtained for such methods are better than those available for the Lasso and related ℓ_1-penalized techniques. Furthermore, the procedure of exponential weighting accomplishes the task of universal aggregation.

We consider the Gaussian regression model with fixed design. Suppose that we observe $\{(X_i, Y_i)\}_{i=1}^n$ such that

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \dots, n, \tag{1.1}$$

where $\mathcal{X}$ is an arbitrary set, $f : \mathcal{X} \to \mathbb{R}$ is an unknown function, $X_i \in \mathcal{X}$ are nonrandom, and ξ_i are independent random variables. Unless explicitly stated otherwise, we will assume

Proceedings of the International Congress of Mathematicians, Seoul, 2014

that ξ_i are independent identically distributed (i.i.d.) Gaussian random variables with mean zero and variance σ^2, $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

Let $\hat{f}$ be an estimator of f based on the observations $\{(X_i, Y_i)\}_{i=1}^n$. To measure the performance of $\hat{f}$, we use the squared error loss of the form

$$\|\hat{f} - f\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f(X_i))^2$$

and we define the squared risk of estimator $\hat{f}$ as $E\|\hat{f} - f\|^2$ where E denotes the expectation sign. The pseudo-norm $\|f\|$ is referred to as the *empirical norm* of a function defined on $\mathcal{X}$. For vectors $b \in \mathbb{R}^n$, we will also consider the empirical ℓ_2-norm defined by $\|b\|^2 = \frac{1}{n}\sum_{i=1}^n b_i^2$, while $|b|_2^2 = \sum_{i=1}^n b_i^2$ defines the usual ℓ_2-norm $|b|_2$.

Assume that we are given a collection of functions $\{f_1, \dots, f_M\}$ called the *dictionary*, where $f_j : \mathcal{X} \to \mathbb{R}$. Assume also that we are given a subset Θ of $\mathbb{R}^M$. For $\theta = (\theta_1, \dots, \theta_M) \in \Theta$ we consider the linear combinations f_θ defined by

$$\mathsf{f}_\theta(x) \stackrel{\text{def}}{=} \sum_{j=1}^{M} \theta_j f_j(x), \quad x \in \mathcal{X}.$$

Functions f_θ are viewed as approximations of the unknown f. Choosing the dictionary $\{f_1, \dots, f_M\}$ to be rich enough and M sufficiently large, one can expect f_θ to be close to f under appropriate assumptions. This motivates the study of estimator of f having the form

$$\hat{f} = \mathsf{f}_{\hat{\theta}} = \sum_{j=1}^{M} \hat{\theta}_j f_j,$$

where $\hat{\theta}_j$ are suitable estimators of the coefficients θ_j. The overall aim is to mimimize the risk by choosing an optimal $\hat{\theta}_j$. However, depending on the assumptions that we make about the dictionary, the set Θ and f, we are led to different optimality properties. We introduce here three different settings and discuss the corresponding minimax optimality frameworks.

1.1. Setting 1: Linear regression and sparsity. Assume that f is a linear combination of functions from the dictionary:

$$\exists \theta^* \in \mathbb{R}^M : \quad f(x) = \mathsf{f}_{\theta^*}(x) = \sum_{j=1}^{M} \theta_j^* f_j(x). \tag{1.2}$$

Then (1.1) takes the form of a linear regression model, i.e., it can be written as

$$y = X\theta^* + \xi, \tag{1.3}$$

where $y = (Y_1, \dots, Y_n)^T$, $\xi = (\xi_1, \dots, \xi_n)^T$, and $X \in \mathbb{R}^{n \times M}$ is a deterministic matrix with entries $f_j(X_i), i = 1, \dots, n, j = 1, \dots, M$. Estimation of f is now reduced to estimation of θ^* in (1.3). Classical theory of linear regression deals with the case $n \le M$, which is a necessary condition of identifiability of θ^* when we only know that $\theta^* \in \mathbb{R}^M$. However, motivated by several applications, recent years have witnessed an increasing interest in the

problems where M is greater than n and often $M \gg n$. In this case, f is not identifiable without additional assumptions on θ^*. A natural and most popular additional assumption is a sparsity constraint. It consists in restricting the parameter θ^* to the class $\Theta = B_0(s)$ where $B_0(s)$ is the ℓ_0-ball in $\mathbb{R}^M$:

$$B_0(s) = \{\theta \in \mathbb{R}^M : |\theta|_0 \le s\}, \quad s = 1, \dots, M. \tag{1.4}$$

Here,

$$|\theta|_0 \stackrel{\text{def}}{=} \sum_{j=1}^{M} I(\theta_j \neq 0)$$

is the "ℓ_0 norm" of θ, where $I(\cdot)$ denotes the indicator function. Vectors θ belonging to $B_0(s)$ are called s-sparse. It turns out that, under the s-sparisty restriction, estimation with reasonable accuracy is possible. A natural question arising in this context is: What is the optimal way to estimate θ^* if we know that $\theta^* \in B_0(s)$?

We will consider optimality in a minimax sense. Let $\hat{\theta}$ be an estimator of θ^*. The corresponding estimator of f is then $\hat{f} = \mathsf{f}_{\hat{\theta}}$ and, in view of (1.2) - (1.3), the squared risk takes the form

$$E\|\hat{f} - f\|^2 = E_{\theta^*}\left(\frac{1}{n}|X(\hat{\theta} - \theta^*)|_2^2\right).$$

Here and below, E_θ denotes the expectation with respect to the distribution of $y = X\theta + \xi$ where ξ is a Gaussian vector in $\mathbb{R}^n$ with i.i.d. components $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

An estimator $\hat{\theta}$ is called minimax optimal on the class $B_0(s)$ if there exists a sequence of positive numbers $\psi_{n,M,s}$ such that, for all n and M, the following two conditions are satisfied:

$$\sup_{\theta^* \in B_0(s)} E_{\theta^*}\left(\frac{1}{n}|X(\hat{\theta} - \theta^*)|_2^2\right) \le C\psi_{n,M,s}, \tag{1.5}$$

$$\inf_{T} \sup_{\theta^* \in B_0(s)} E_{\theta^*}\left(\frac{1}{n}|X(T - \theta^*)|_2^2\right) \ge c\psi_{n,M,s} \tag{1.6}$$

where C and c are positive constants independent of n, M, s, and $\inf_T$ denotes the infimum over all estimators of θ^* based on the sample $\{(X_i, Y_i)\}_{i=1}^n$ satisfying the model (1.3). This property is commonly referred to as the *minimax optimality*. A sequence $\psi_{n,M,s}$ such that (1.5) and (1.6) hold is called *minimax rate of convergence* (or *optimal rate of convergence*) on $B_0(s)$. Our main goal in this setting is to find a minimax optimal estimator $\hat{\theta}$ on the class $B_0(s)$. Along with $B_0(s)$, other classes can be considered, such as the ℓ_q-balls $B_q(\delta) = \{\theta \in \mathbb{R}^M : |\theta|_q \le \delta\}$ where $|\theta|_q = (\sum_{j=1}^M |\theta_j|^q)^{\frac{1}{q}}$, $0 < q < \infty$, $\delta > 0$. The notion of optimality is defined for them analogously. This problem, in its particular case where $X^T X/n$ is the identity matrix, and $M = n$ (called the Gaussian sequence model) and with an asymptotic point of view ($n \to \infty$), has been in the focus of the statistical literature since the 1990ies [1, 16, 17]; for its non-asymptotic treatment, see [4]. We are interested here in a more general linear regression setting and we deal with the non-asymptotic minimax optimality. We also consider more general classes, such as intersection of ℓ_0-ball with ℓ_q-ball, $0 < q \le 2$, for which we propose an adaptive estimator. Here, adaptivity means that the estimator is independent of s, q, and of the radius δ of the ℓ_q-ball and it achieves the minimax rates simultateously for all $1 \le s \le M$, $\delta > 0$, $0 < q \le 2$.

1.2. Setting 2: Nonparametric regression. Let $\mathcal{F}_{\beta,L}$ be a class of smooth functions on $\mathcal{X} \subseteq \mathbb{R}^d$ indexed by $\beta > 0$ and $L > 0$. Common examples of $\mathcal{F}_{\beta,L}$ are balls in a Sobolev or Besov space (see [19, 34] for more details). Assume that $f \in \mathcal{F}_{\beta,L}$. The parameter β typically stands for the number of derivatives of f that are assumed bounded in some norm by L, the radius of the ball. In this setting, the dictionary $\{f_1, \dots, f_M\}$ is usually composed of the first $M = n$ functions of some orthonormal basis. For example, it can be the Fourier or wavelet basis. A key property in the nonparametric regression setting (following from the definition of the class $\mathcal{F}_{\beta,L}$) is that f can be well approximated by a linear combination of basis functions. It can be stated, for example, as follows: For any $f \in \mathcal{F}_{\beta,L}$, and any $k = 1, \dots, n$, there exists $\theta^* = \theta^*(f) \in \mathbb{R}^k$ such that

$$\Big\| f - \sum_{j=1}^{k} \theta_j^* f_j \Big\| \le C(\beta, L) k^{-\beta}, \tag{1.7}$$

where $C(\beta, L)$ is a constant depending only on β, L. A minimax optimal estimator $\hat{f}$ is the estimator that satisfies, for all n,

$$\sup_{f \in \mathcal{F}_{\beta,L}} E\|\hat{f} - f\|^2 \le C\psi_{n,\beta}, \tag{1.8}$$

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{F}_{\beta,L}} E\|\tilde{f} - f\|^2 \ge c\psi_{n,\beta}, \tag{1.9}$$

where C and c are positive constants independent of n, β, and $\inf_{\tilde{f}}$ denotes the infimum over all estimators of f based on the sample $\{(X_i, Y_i)\}_{i=1}^n$. A positive sequence $\psi_{n,\beta}$ such that (1.8) and (1.9) hold is called the *minimax rate of convergence* (or *optimal rate of convergence*) on $\mathcal{F}_{\beta,L}$.

Along with finding minimax optimal estimators in this setting, the second important issue is adaptivity: How to construct *adaptive estimators* that is estimators $\hat{f}$, which are independent of β and L and satisfy (1.8) with optimal rate of convergence $\psi_{n,\beta}$ for all pairs (β, L) in a wide range of values?

1.3. Setting 3: Aggregation of estimators. This setting will be the main object of study below. Suppose that we are given a collection of estimators $\hat{f}_1, \dots, \hat{f}_M$ of f and a subset Θ of $\mathbb{R}^M$. The goal is to find a new estimator $\tilde{f}$, called the *aggregate*, which is approximately at least as good as the best linear combination $\mathsf{f}_\theta = \sum_{j=1}^M \theta_j \hat{f}_j$ with weights θ in the set Θ. The best linear combination is defined as the one that solves the minimization problem

$$\min_{\theta \in \Theta} E\|f - \mathsf{f}_\theta\|^2.$$

Unlike in the previous two settings, here f_θ is a random function depending on the data, and we *do not assume* that $\|f - \mathsf{f}_\theta\|$ is zero or small (cf. (1.2), (1.7)); it may happen that all f_θ for some Θ are very far from f. Some common examples of Θ are the following.

(L) *Linear aggregation:* $\Theta = \mathbb{R}^M$. The aim of linear aggregation is to construct an estimator $\tilde{f}$, which is approximately as good as the best linear combination of the initial estimators $\hat{f}_1, \dots, \hat{f}_M$.

(C) *Convex aggregation:* Θ is the simplex

$$\Theta = \Lambda^M \stackrel{\text{def}}{=} \Big\{ \theta \in \mathbb{R}^M : \theta_j \ge 0, \sum_{j=1}^M \theta_j = 1 \Big\}.$$

Convex aggregation aims to find an estimator $\tilde{f}$, which is approximately as good as the best convex combination of the initial estimators $\hat{f}_1, \dots, \hat{f}_M$.

(MS) *Model Selection type aggregation:* $\Theta = \{e_1, \dots, e_M\}$ where e_i are the canonical basis vectors in $\mathbb{R}^M$. The MS-aggregation aims to mimic the best among the initial estimators $\hat{f}_1, \dots, \hat{f}_M$.

Other types of aggregation can be considered as well, for example, the *s-sparse aggregation* corresponding to $\Theta = B_0(s)$, or the ℓ_q-aggregation corresponding to $\Theta = B_q(\delta)$ with $0 < q < \infty$, $\delta > 0$.

The goal of aggregation is to mimic the best linear combination of initial estimators with weights restricted to a given set Θ of possible weights. The word "best" here is formalized as choosing $\tilde{f}$ with the smallest possible *excess risk* (also known under the name of *regret*) defined by

$$\mathcal{E}_\Theta(\tilde{f}, f) \stackrel{\text{def}}{=} E\|\tilde{f} - f\|^2 - \inf_{\theta \in \Theta} E\|\mathsf{f}_\theta - f\|^2. \tag{1.10}$$

Based on the excess risk, we can introduce the concept of minimax optimality for aggregation [33]. An estimator $\tilde{f}$ is called an *optimal aggregate for the class* Θ if there exists a sequence of positive numbers $\psi_{n,M}(\Theta)$ such that, for all n and M,

$$\sup_{\hat{f}_1, \dots, \hat{f}_M} \big\{ \sup_f \mathcal{E}_\Theta(\tilde{f}, f) \big\} \le C\psi_{n,M}(\Theta), \tag{1.11}$$

$$\sup_{\hat{f}_1, \dots, \hat{f}_M} \big\{ \inf_{\hat{f}} \sup_f \mathcal{E}_\Theta(\hat{f}, f) \big\} \ge c\psi_{n,M}(\Theta). \tag{1.12}$$

Here, $\inf_{\hat{f}}$ is the minimum over all estimators, C and c are positive constants independent of n and M, and $\sup_f$, $\sup_{\hat{f}_1, \dots, \hat{f}_M}$ are the suprema over all possible functions f and over wide classes of preliminary estimators. In some cases, these will be all possible estimators $\hat{f}_1, \dots, \hat{f}_M$ with no restriction; in other cases it will suffice to consider classes of $\hat{f}_1, \dots, \hat{f}_M$ such that $\hat{f}_j$'s are bounded in the empirical norm $\|\cdot\|$ uniformly over j. If (1.11) and (1.12) hold for some positive sequence $\psi_{n,M}(\Theta)$, this sequence is called an *optimal rate of aggregation for the class* Θ [33]. Two questions arise in this context. First, how to construct an optimal aggregate $\tilde{f}$ for a given class Θ? Second, is it possible to construct a *universal aggregate*, i.e., an aggregate which is optimal simultaneously for a large scale of classes Θ?

Inequalities (1.11) and (1.12) establish upper and lower bounds for the minimax regret, respectively. The upper bound (1.11) can be equivalently written in the form of *oracle inequality*[1]

$$E\|\tilde{f} - f\|^2 \le \inf_{\theta \in \Theta} E\|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta), \tag{1.13}$$

that should be valid for all $\hat{f}_1, \dots, \hat{f}_M$ in a wide class, and for all f. This garantees that the risk of the suggested aggregate $\tilde{f}$ is at least as good as the risk of the unknown *oracle* θ^* minimizing $E\|\mathsf{f}_\theta - f\|^2$, up to a remainder term of the order $\psi_{n,M}(\Theta)$, which characterizes the price to pay for aggregation. Lower bounds (1.12) ensure that this is the minimal price; the remainder term cannot be of smaller order whatever the aggregate is. For sparsity classes, for example, when $\Theta = B_0(s)$, the rate $\psi_{n,M}(\Theta)$ is a function of s; the corresponding oracle inequalities are called *sparsity oracle inequalities*.

[1]Here and in the sequel, we denote by C positive constants, possibly different on different appearances.

2. Reduction to aggregation of functions

Aggregates are usually constructed in the form

$$\tilde{f} = \sum_{j=1}^{M} \hat{\theta}_j \hat{f}_j$$

where $\hat{\theta}_j$ are suitably chosen statistics measurable with respect to the data, and $\hat{f}_j$ are the preliminary estimators. In what follows, we will assume that $\hat{\theta}_j$ and estimators $\hat{f}_j$ are stochastically independent. This can be achieved by creating two independent samples from the initial sample $\{(X_i, Y_i)\}_{i=1}^n$ by randomization (*sample cloning*), cf. [27]. The estimators $\hat{f}_j$ are constructed from the first sample while the second one is used to perform aggregation, i.e., to compute the weights $\hat{\theta}_j$. To carry out the analysis of aggregation, it is enough to work conditionally on the first sample, so that $\hat{f}_j$ can be considered as deterministic functions. Thus, the problem reduces to aggregation of deterministic functions that we will denote as previously $f_j \stackrel{\text{def}}{=} \hat{f}_j$, $j = 1, \dots, M$. The procedure of sample cloning by randomization is based on the following elementary fact.

Lemma 2.1. *Let $Y_i = f(X_i) + \xi_i$. Let ω_i be a standard normal random variable independent of ξ_i. Set $Y_{i1} = Y_i + \sigma\omega_i$, and $Y_{i2} = Y_i - \sigma\omega_i$. Then we have $Y_{i1} = f(X_i) + \xi_{i1}$, and $Y_{i2} = f(X_i) + \xi_{i2}$, where $\xi_{i1} \sim \mathcal{N}(0, 2\sigma^2)$, $\xi_{i2} \sim \mathcal{N}(0, 2\sigma^2)$ and ξ_{i1} is independent of ξ_{i2}.*

Thus, by adding to and subtracting from the observations Y_i the variables $\sigma\omega_i$, we obtain two independent Gaussian n-samples $D_1 = \{(X_i, Y_{i1})\}_{i=1}^n$ and $D_2 = \{(X_i, Y_{i2})\}_{i=1}^n$, where $Y_{ik} = f(X_i) + \xi_{ik}$, $k = 1, 2$. The observations in both samples are of the same form as in the original sample $\{(X_i, Y_i)\}_{i=1}^n$, with the only difference that the variance of the noise is doubled.

Now, we use D_1 to construct preliminary estimators $\hat{f}_1, \dots, \hat{f}_M$ and we use D_2 to determine the weights $\hat{\theta}_1, \dots, \hat{\theta}_M$. Denoting by $E_{(k)}$ the expectations with respect to the distribution of D_k for $k = 1, 2$, we may write the oracle inequality (1.13) that we need to prove in the form

$$E_{(1)}E_{(2)}\|\tilde{f} - f\|^2 \le \inf_{\theta\in\Theta} E_{(1)}\|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta). \tag{2.1}$$

To obtain (2.1), it suffices to show that, for fixed functions $f_1, \dots, f_M, f$, we have

$$E_{(2)}\|\tilde{f} - f\|^2 \le \inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta), \tag{2.2}$$

where $\mathsf{f}_\theta = \sum_{j=1}^M \theta_j f_j$, and $\tilde{f} = \sum_{j=1}^M \hat{\theta}_j f_j$ with $\hat{\theta}_j$ measurable with respect to D_2.

Thus, using the sample cloning device, we can reduce aggregation of estimators to its special case, which is *aggregation of fixed functions*. This will be the setting considered in the rest of the paper. In this case, the minimax framework of aggregation (cf. Setting 3 in the Introduction) changes only in that the excess risk takes the form

$$\mathcal{E}_\Theta(\tilde{f}, f) \stackrel{\text{def}}{=} E\|\tilde{f} - f\|^2 - \inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 \tag{2.3}$$

(no expectation in the term $\inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2$). Accordingly, an estimator $\tilde{f}$ is called an *optimal aggregate for the class* Θ if there exists a sequence of positive numbers $\psi_{n,M}(\Theta)$

such that (1.11) and (1.12) are satisfied where instead of $\hat{f}_j$ we have fixed functions f_j. Upper bounds on the maximum excess risk are then equivalent to oracle inequalities

$$E\|\tilde{f} - f\|^2 \le \inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta). \tag{2.4}$$

Such an oracle inequality being established, we can obtain upper bounds for the minimax risks in Settings 1 and 2 as simple corollaries. Indeed, those settings impose additional strong restrictions on f; the oracle risk $\inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2$ is 0 in Setting 1, and it admits a bound such as (1.7) in Setting 2. In Setting 3, the oracle risk can be arbitrary, therefore we use only the excess risk to measure the performance of aggregates.

3. Least squares aggregation

A first simple idea is to construct aggregates by minimizing the least squares (LS) criterion. Given a set Θ and a collection of deterministic functions $f_1, \dots, f_M$, we take

$$\hat{\theta}^{LS}(\Theta) \in \operatorname*{argmin}_{\theta\in\Theta} \|y - \mathsf{f}_\theta\|^2$$

and we define the LS aggregate as

$$\tilde{f} = \mathsf{f}_{\hat{\theta}^{LS}(\Theta)} = \sum_{j=1}^{M} \hat{\theta}_j^{LS}(\Theta) f_j.$$

Proposition 3.1. *Let $\hat{\theta}^{LS} \stackrel{\text{def}}{=} \hat{\theta}^{LS}(\mathbb{R}^M)$ be a global least squares estimator. Assume that $E(\xi_i) = 0$, $E(\xi_i\xi_j) = 0$, if $i \ne j$ for $i, j = 1, \dots, n$. If $E(\xi_i^2) = \sigma^2$, $i = 1, \dots, n$, then for all $f, f_1, \dots, f_M$, and integers $n \ge 1$, $M \ge 1$, we have*

$$E\|\mathsf{f}_{\hat{\theta}^{LS}} - f\|^2 = \min_{\theta\in\mathbb{R}^M} \|\mathsf{f}_\theta - f\|^2 + \frac{\sigma^2 R}{n}. \tag{3.1}$$

where $R = \operatorname{Rank}(X)$ denotes the rank of matrix X. Furthermore, if $E(\xi_i^2) \le \sigma^2$, $i = 1, \dots, n$, then under the same other assumptions for any convex set $\Theta \subset \mathbb{R}^M$,

$$E\|\mathsf{f}_{\hat{\theta}^{LS}(\Theta)} - f\|^2 \le \min_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + \frac{4\sigma^2 R}{n}. \tag{3.2}$$

Proof. We prove only (3.2) since (3.1) follows from a simple orthogonal decomposition, cf., e.g., [31]. Below, we will denote by f and f_θ not only the functions from $\mathcal{X}$ to $\mathbb{R}$ but also the n-vectors of values of these functions at points $X_1, \dots, X_n$. Then, the model of observations (1.1) can be written as $y = f + \xi$, and $\mathsf{f}_\theta = X\theta$ for all θ. Set for brevity $\tilde{f} = \mathsf{f}_{\hat{\theta}^{LS}(\Theta)}$. From the definition of this estimator we get by a simple algebra that, for any $\theta \in \Theta$,

$$\|\tilde{f} - f\|^2 \le \|\mathsf{f}_\theta - f\|^2 + 2\langle \tilde{f} - \mathsf{f}_\theta, \xi\rangle \tag{3.3}$$

where $\langle f, g\rangle \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)$. On the other hand, $\tilde{f} - \mathsf{f}_\theta \in \operatorname{Im}(X)$, and thus $\langle \tilde{f} - \mathsf{f}_\theta, \xi\rangle = \langle \tilde{f} - \mathsf{f}_\theta, A\xi\rangle$ where A is the orthogonal projector on $\operatorname{Im}(X)$. This and (3.3) imply

$$\|\tilde{f} - f\|^2 \le \|\mathsf{f}_\theta - f\|^2 + \frac{1}{2}\|\tilde{f} - \mathsf{f}_\theta\|^2 + 2\|A\xi\|^2. \tag{3.4}$$

Since Θ is convex, for all $\theta' \in \Theta$ we have $\|f - \mathsf{f}_{\theta^*}\|^2 + \|\mathsf{f}_{\theta'} - \mathsf{f}_{\theta^*}\|^2 \le \|\mathsf{f}_{\theta'} - f\|^2$ where $\theta^* = \operatorname*{argmin}_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2$. This inequality with $\theta' = \hat\theta^{LS}(\Theta)$, and (3.4) with $\theta = \theta^*$ imply that $\|\tilde f - f\|^2 \le \|\mathsf{f}_{\theta^*} - f\|^2 + 4\|A\xi\|^2$. Now, (3.2) follows by taking here the expectations and noticing that $E\|A\xi\|^2 \le \frac{\sigma^2 R}{n}$. □

Proposition 3.2. *Let Θ be a subset of the simplex Λ^M, and let $\xi_1, \dots, \xi_n$ be independent zero mean σ-subgaussian random variables, i.e., $E\exp(s\xi_i) \le \exp(s^2\sigma^2/2)$ for all $s > 0$, $i = 1, \dots, n$. Then, for all f, all integers $n \ge 1$, $M \ge 2$, and all dictionaries $\{f_1, \dots, f_M\}$ such that $\max_{j=1,\dots,M} \|f_j\| \le L$, we have*

$$E\|\mathsf{f}_{\hat\theta^{LS}(\Theta)} - f\|^2 \le \inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + 2\sigma L\sqrt{\frac{2\log M}{n}}\,.$$

Proof. In view of (3.3), it suffices to prove that $E\langle \tilde f, \xi\rangle \le \sigma L\sqrt{\frac{2\log M}{n}}$ where $\tilde f = \mathsf{f}_{\hat\theta^{LS}(\Theta)}$. But $E\langle \tilde f, \xi\rangle \le E\max_{\theta'\in\Lambda^M}\langle f_{\theta'}, \xi\rangle = E\max_{1\le j\le M}\langle f_j, \xi\rangle$ and the random variable $\langle f_j, \xi\rangle$ is $\bar\sigma$-subgaussian with $\bar\sigma = \sigma\|f_j\|/\sqrt{n} \le \sigma L/\sqrt{n}$. By the standard properties of subgaussian variables, $E\max_{1\le j\le M}\langle f_j, \xi\rangle \le \bar\sigma\sqrt{2\log M}$. □

Consider now convex aggregation and MS-aggregation by the LS method. The corresponding weights are $\hat\theta^{LS}_{\mathrm{conv}} \stackrel{\mathrm{def}}{=} \hat\theta^{LS}(\Lambda^M)$, and $\hat\theta^{LS}_{\mathrm{MS}} \stackrel{\mathrm{def}}{=} \hat\theta^{LS}(\{e_1, \dots, e_M\})$. The MS-aggregate selects one function in the dictionary:

$$\mathsf{f}_{\hat\theta^{LS}_{\mathrm{MS}}} = f_{\hat\jmath} \quad \text{where} \quad \hat\jmath \in \operatorname{argmin}_{1\le j\le M}\|y - f_j\|^2.$$

The following corollaries are straightforward in view of Propositions 3.1 and 3.2.

Corollary 3.3 (Convex aggregation). *For all f, all integers $n \ge 1$, $M \ge 2$, and all dictionaries $\{f_1, \dots, f_M\}$ such that $\max_{j=1,\dots,M} \|f_j\| \le L$, we have*

$$E\|\mathsf{f}_{\hat\theta^{LS}_{\mathrm{conv}}} - f\|^2 \le \min_{\theta\in\Lambda^M} \|\mathsf{f}_\theta - f\|^2 + \left(\frac{4\sigma^2 R}{n} \wedge 2\sigma L\sqrt{\frac{2\log M}{n}}\right).$$

Corollary 3.4 (MS-aggregation). *For all f, all integers $n \ge 1$, $M \ge 2$, and all dictionaries $\{f_1, \dots, f_M\}$ such that $\max_{j=1,\dots,M} \|f_j\| \le L$, we have*

$$E\|\mathsf{f}_{\hat\theta^{LS}_{\mathrm{MS}}} - f\|^2 \le \min_{1\le j\le M} \|f_j - f\|^2 + 2\sigma L\sqrt{\frac{2\log M}{n}}\,.$$

The rate of aggregation $\frac{\sigma^2 R}{n}$ of the global least squares estimator given in (3.1) is the optimal rate of linear aggregation, see Section 8 below and [6, 31, 33]. Also, the rate of the convex aggregate $\mathsf{f}_{\hat\theta^{LS}_{\mathrm{conv}}}$ given in Corollary (3.3) is the optimal rate of convex aggregation up to a minor discrepancy in the expression under the logarithm [31, 33]. However, for MS-aggregation the situation is different. The optimal rate for MS-aggregation is of the order $(\log M)/n$ [31, 33], while the LS-aggregate $\mathsf{f}_{\hat\theta^{LS}_{\mathrm{MS}}}$ achieves only the rate $\sqrt{(\log M)/n}$ according to Corollary 3.4. Moreover, it turns out that $\mathsf{f}_{\hat\theta^{LS}_{\mathrm{MS}}}$ cannot do better; the upper

bound of Corollary 3.4 is tight for it. Indeed, the next theorem shows that not only the least squares MS-aggregate but also any method that selects a single function in the dictionary cannot have faster rate. This includes methods of model selection by penalized empirical risk minimization. We call estimators $\hat{S}_n$ taking values in $\{f_1, \dots, f_M\}$ the *selectors*.

Theorem 3.5 ([32]). *Assume that $n \geq 1$, $M \geq 2$ are such that*

$$(\sigma \vee 1)\sqrt{(\log M)/n} \leq C_0$$

for $0 < C_0 < 1$ small enough. Then, there exists a dictionary $\{f_1, \dots, f_M\}$ with $\|f_j\| \leq 1$, $j = 1, \dots, M$, such that the following holds. For any selector $\hat{S}_n$, there exists a regression function f such that $\|f\| \leq 1$ and

$$E\|\hat{S}_n - f\|^2 \geq \min_{1 \leq j \leq M} \|f_j - f\|^2 + C_* \sigma \sqrt{\frac{\log M}{n}}$$

for some positive constant C_* independent of n and M.

A related result about suboptimality of selectors when $\hat{f}_j$ are preliminary estimators rather than fixed functions is proved in [18].

Thus, we see that choosing one of the functions in a finite dictionary to solve the problem of model selection is suboptimal in the sense that the rate $\sqrt{(\log M)/n}$ is too slow. A natural idea is to extend the class of estimators by taking a convex combination of the functions in the dictionary rather than selecting one function. It turns out that this is sufficient; under a particular choice of weights in this convex combination, namely the exponential weights, one can achieve oracle inequalities with the optimal rate $(\log M)/n$.

4. Exponentially weighted aggregates

Let $f_1, \dots, f_M$ be a given dictionary of functions. Consider the exponentially weighted aggregate

$$\hat{f}^{EW} \stackrel{\text{def}}{=} \mathsf{f}_{\hat{\theta}^{EW}} = \sum_{j=1}^{M} \hat{\theta}_j^{EW} f_j$$

where the weights $\hat{\theta}^{EW} = (\hat{\theta}_1^{EW}, \dots, \hat{\theta}_M^{EW})$ are defined as

$$\hat{\theta}_j^{EW} = \frac{\exp(-n\hat{r}_j/\beta)\pi_j}{\sum_{k=1}^{M} \exp(-n\hat{r}_k/\beta)\pi_k}\,.$$

Here, $\hat{r}_j = \|y - f_j\|^2$ is the empirical risk corresponding to function f_j, $\beta > 0$ is a tuning parameter, and $\pi_1, \dots, \pi_M$ is a set of prior probabilities, $\pi_k > 0$, $\sum_{k=1}^{M} \pi_k = 1$. This definition dates back at least to [37] where the method was introduced in the context of the theory of prediction of deterministic individual sequences. It is now a popular tool in that theory, cf. [8, 18] where one can find further references.

Note that

$$\hat{\theta}^{EW} = \operatorname*{argmin}_{\theta \in \Lambda^M} \left(\sum_{j=1}^{M} \theta_j \hat{r}_j + \frac{\beta}{n} \mathcal{K}(\theta, \pi) \right) \tag{4.1}$$

where $\mathcal{K}(\theta,\pi)=\sum_{j=1}^M \theta_j \log\frac{\theta_j}{\pi_j}\ge 0$ (with the convention that $0\cdot\log 0=0$) is the Kullback-Leibler divergence between the discrete probability measures defined by the probability vectors $\theta\in\Lambda^M$ and $\pi\in\Lambda^M$. Since, by Jensen's inequality, $\sum_{j=1}^M \theta_j\hat r_j\ge\|y-\mathsf{f}_\theta\|^2$, we see that $\hat\theta^{EW}$ minimizes, over the simplex Λ^M, an upper bound on the empirical risk penalized by the Kullback-Leibler divergence from π:

$$\|y-\mathsf{f}_\theta\|^2+\frac{\beta}{n}\mathcal{K}(\theta,\pi).$$

So, intuitively, the method penalizes the solution for being far from the prior π.

Theorem 4.1. *For $\beta\ge 4\sigma^2$, and for all $f, f_1,\dots,f_M$, and integers $n\ge 1$, $M\ge 1$, we have*

$$E\|\hat f^{EW}-f\|^2\le\min_{\theta\in\Lambda^M}\left(\sum_{j=1}^M\theta_j\|f-f_j\|^2+\frac{\beta}{n}\mathcal{K}(\theta,\pi)\right). \tag{4.2}$$

If the ξ_i are not Gaussian but rather i.i.d. symmetric random variables such that $P(|\xi_i|\le B)=1$ for some finite $B>0$, then (4.2) *holds for any $\beta\ge 4B^2$.*

The proof of this theorem can be found in [12, 13, 15] as a special case of more general results relaxing the assumptions on the distribution of ξ_i and allowing for continuous priors (see also [14]). More recent work [9, 23, 29] proposes estimators other than $\hat f^{EW}$ satisfying analogous oracle inequalities both in expectation and in probability.

Note that the right-hand side of (4.2) is similar to (4.1). The only difference is that in (4.1) we have the empirical risks $\hat r_j=\|y-f_j\|^2$ rather than the deterministic discrepancies $\|f-f_j\|^2$. Thus, the minimization problem in (4.1) is an empirical analog of the right-hand side of (4.2). An immediate corollary of Theorem 4.1 is the following.

Theorem 4.2. *For $\beta\ge 4\sigma^2$, and for all $f, f_1,\dots,f_M$, and integers $n\ge 1$, $M\ge 1$, we have*

$$E\|\hat f^{EW}-f\|^2\le\min_{1\le j\le M}\left(\|f-f_j\|^2+\frac{\beta}{n}\log\frac{1}{\pi_j}\right).$$

In particular, if $\pi_j=\frac{1}{M}$, $j=1,\dots,M$, and $M\ge 2$,

$$E\|\hat f^{EW}-f\|^2\le\min_{1\le j\le M}\|f-f_j\|^2+\frac{\beta}{n}\log M.$$

Thus, the exponentially weighted aggregate achieves the optimal rate of the order $(\log M)/n$, which cannot be attained by the selectors. A result similar to Theorem 4.2 was proved in [24] for the case where f_j are not arbitrary fixed functions but rather the least squares estimators on linear subspaces of $\mathbb{R}^M$. In [24], these estimators are constructed from the same sample y that is used to compute the weights, and the weights are are different:

$$w_j=\frac{\exp\left(-\frac{n\hat r_j}{\beta}-\frac{\dim(j)}{2}\right)\pi_j}{\sum_{k=1}^M\exp\left(-\frac{n\hat r_k}{\beta}-\frac{\dim(k)}{2}\right)\pi_k} \tag{4.3}$$

where $\dim(j)$ is the dimension of the space on which the jth least squares estimator projects. Extension of the results of [24] to affine estimators are given in [10, 11].

5. Sparsity pattern aggregation

We call a *sparsity pattern* any binary vector $p \in \mathcal{P} \stackrel{\text{def}}{=} \{0,1\}^M$. Denote by $|p| = |p|_0$ the number of ones in p. To each sparsity pattern $p = (p_1, \ldots, p_M) \in \mathcal{P}$, we associate a linear subspace S^p of $\mathbb{R}^M$:

$$S^p \stackrel{\text{def}}{=} \operatorname{span}\{e_j : p_j = 1\}, \qquad \dim(S^p) = |p|.$$

From the initial sample y, we clone two randomized independent samples $y^{(1)} \in \mathbb{R}^n$ and $y^{(2)} \in \mathbb{R}^n$ with $\mathcal{N}(0, 2\sigma^2)$ errors as described in Section 2. For each $p \in \mathcal{P}$, we construct a least squares estimator $\hat{\theta}_p$ on S^p based on the first sample $y^{(1)}$:

$$\hat{\theta}_p \in \operatorname*{argmin}_{\theta \in S^p} \|y^{(1)} - \mathsf{f}_\theta\|^2. \tag{5.1}$$

Set $\hat{r}_p = \|y^{(2)} - \mathsf{f}_{\hat{\theta}_p}\|^2$ and define a vector $\hat{\theta}^{SPA} = (\hat{\theta}^{SPA}_p,\ p \in \mathcal{P})$ with components

$$\hat{\theta}^{SPA}_p = \frac{\exp(-n\hat{r}_p/\beta)\pi_p}{\sum_{p' \in \mathcal{P}} \exp(-n\hat{r}_{p'}/\beta)\pi_{p'}}, \quad \forall p \in \mathcal{P}.$$

Here, $\{\pi_p, p \in \mathcal{P}\}$ is a prior probability measure on $\mathcal{P}$ with $\pi_p \geq 0$ (not necessarily $\pi_p > 0$; values $\pi_p = 0$ are possible, as opposed to the priors in Section 4). The *sparsity pattern aggregate* is defined by

$$\hat{f}^{SPA} \stackrel{\text{def}}{=} \sum_{p \in \mathcal{P}} \hat{\theta}^{SPA}_p \mathsf{f}_{\hat{\theta}_p}.$$

From Theorem 4.2 (where we replace σ^2 by $2\sigma^2$ to account for the sample cloning) we get that if $\beta = 8\sigma^2$, then

$$\forall f: \quad E\|\hat{f}^{SPA} - f\|^2 \leq \min_{p \in \mathcal{P}: \pi_p \neq 0} \left[E\|f_{\hat{\theta}_p} - f\|^2 + \frac{8\sigma^2}{n} \log \frac{1}{\pi_p} \right] \tag{5.2}$$

while from Proposition 3.1 (again, replacing σ^2 by $2\sigma^2$),

$$\forall f: \quad E\|\mathsf{f}_{\hat{\theta}_p} - f\|^2 \leq \min_{\theta \in S^p} \|\mathsf{f}_\theta - f\|^2 + \frac{2\sigma^2|p|}{n}. \tag{5.3}$$

Consider the prior distribution

$$\pi_p = \begin{cases} \left(\binom{M}{|p|} e^{|p|} H\right)^{-1} & \text{if } |p| \leq R, \\ 1/2 & \text{if } |p| = M, \\ 0 & \text{otherwise,} \end{cases} \tag{5.4}$$

where $H > 0$ is the normalizing constant such that $\sum_{p \in \mathcal{P}} \pi_p = 1$.

Denote by $\hat{f}^{ES}$ the sparsity pattern aggregate $\hat{f}^{SPA}$ with the prior π_p given in (5.4). Following [31], we will call $\hat{f}^{ES}$ the *Exponential Screening* (ES) estimator. The corresponding vector of weights is denoted by $\hat{\theta}^{ES}$. Algorithms of computation of this estimator via Markov Chain Monte-Carlo schemes are discussed and analyzed in [31, 32].

Combining (5.2) – (5.4) leads to the following sparsity oracle inequality, cf. [31].

Theorem 5.1. *Let $\hat{f}^{ES}$ be the Exponential Screening estimator with $\beta = 8\sigma^2$. Then for all f, $f_1, \dots, f_M$, and all integers $n \geq 1$, $M \geq 1$, we have*

$$E\|\hat{f}^{ES} - f\|^2 \leq \min_{\theta \in \mathbb{R}^M} \left[\|\mathsf{f}_\theta - f\|^2 + \frac{C\sigma^2}{n} \left(R \wedge |\theta|_0 \log\left(\frac{eM}{|\theta|_0}\right)\right)\right] + \frac{C'\sigma^2}{n} \tag{5.5}$$

where C, C' are absolute positive constants, and $R = \mathrm{Rank}(X)$. Here and in the sequel, $0 \cdot \log \infty = 0$ by convention.

In [31], the oracle inequality (5.5) is proved for a slightly different estimator, with modified weights (4.3) and without sample cloning. A weaker result of this form, not taking into account the rank of X, is given in [3]. Inequalities close to (5.5) that hold with high probability but without accounting for the rank of X are obtained for some estimators different from $\hat{f}^{ES}$ in [10].

Theorem 5.1 is the main result that will allow us to show that one and the same estimator $\hat{f}^{ES}$ achieves the minimax rates of convergence in the three different settings described in the Introduction. Moreover, it achieves these rates adaptively to the parameters of the classes in the first two settings and to the choice of Θ (universal aggregation) in Setting 3. The rest of the paper is devoted to deriving these properties as corollaries of Theorem 5.1.

Remark 5.2. All the results stated below for the estimators $\hat{f}^{ES} = \mathsf{f}_{\hat{\theta}^{ES}}$ and $\hat{\theta}^{ES}$ are also valid for any other estimators $\mathsf{f}_{\hat{\theta}}$ and $\hat{\theta}$ such that (5.5) holds with $\mathsf{f}_{\hat{\theta}}$ in place of $\hat{f}^{ES}$. Indeed, only (5.5) will be used in the subsequent argument.

6. Sparsity oracle inequalities on ℓ_0-balls

Theorem 5.1 and monotonicity of the function $x \mapsto x \log(eM/x)$ imply the following upper bounds.

Theorem 6.1 ([31]). *Let $\hat{f}^{ES}$ be the exponential screening estimator with $\beta = 8\sigma^2$ and let $\hat{\theta}^{ES}$ denote the corresponding vector of weights. Then for all f, $f_1, \dots, f_M$, and all integers $n \geq 1$, $M \geq 1$, $1 \leq s \leq M$,*

$$E\|\hat{f}^{ES} - f\|^2 \leq \min_{\theta \in B_0(s)} \|\mathsf{f}_\theta - f\|^2 + C\sigma^2 \left(\frac{s}{n} \log\left(\frac{eM}{s}\right) \wedge \frac{R}{n}\right) \tag{6.1}$$

where $C > 0$ is an absolute constant. If the model is linear: $y = X\theta + \xi$, then

$$\sup_{\theta \in B_0(s)} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2/n \leq C\sigma^2 \left(\frac{s}{n} \log\left(\frac{eM}{s}\right) \wedge \frac{R}{n}\right) . \tag{6.2}$$

The bounds of Theorem 6.1 are sparsity oracle inequalities. They cannot be improved in a minimax sense, see Section 8 below. Analogous oracle inequalities with leading constant 1 can be also established for the Lasso and related techniques [22] but they need strong assumptions on the dictionary $\{f_1, \dots, f_M\}$ such as the restricted isometry or restricted eigenvalue condition. In contrast to this, the ES estimator satisfies the sparsity oracle inequalities *under no assumption on the dictionary*. Moreover, Theorem 6.1 shows that the ES estimator simultaneously takes advantage of two types of sparsity: small number of non-zero entries of

θ (ℓ_0 norm) and small rank of matrix X. This is not available for the least squares estimators on $B_0(s)$ studied in [28, 36] among others. Note also that the least squares estimators on $B_0(s)$ cannot achieve the excess risk bound (6.1) with leading constant 1 required for the aggregation setting.

7. Estimation on ℓ_q-balls and on intersection of ℓ_0- and ℓ_q-balls

From Theorem 5.1, we can also deduce oracle inequalities and upper bounds on the risk of the estimator $\hat{f}^{ES}$ on ℓ_q-balls with $0 < q \le 2$. They follow from (5.5) using the "Maurey argument" as first noticed in [5, 6] for $q = 1$. The proof for $q = 1$ is based on the next lemma (cf. [5, 6, 31]).

Lemma 7.1. *Let $\|f_j\| \le L$, $j = 1, \dots, M$, and $1 \le m \le M$. Then, for any f and any $\theta \in \mathbb{R}^M$ there exists $\theta' \in \mathbb{R}^M$ such that $|\theta'|_0 \le m$ and*

$$\|\mathsf{f}_{\theta'} - f\|^2 \le \|\mathsf{f}_\theta - f\|^2 + \frac{L^2 |\theta|_1^2}{m}\,.$$

The case $0 < q < 1$ was considered in [10, 35, 38], and the case $1 < q \le 2$ in [35]. Deriving bounds on the risk over ℓ_q-balls with $0 < q < 1$ from (5.5) can be done based on the following extension of Lemma 7.1.

Lemma 7.2. *Let $\|f_j\| \le L$, $j = 1, \dots, M$, and $1 \le m \le M$. Then, for any f, any $0 < q \le 1$ and any $\theta \in \mathbb{R}^M$ there exists $\bar{\theta} \in \mathbb{R}^M$ such that $|\bar{\theta}|_0 \le 2m$ and*

$$\|\mathsf{f}_{\bar{\theta}} - f\|^2 \le \|\mathsf{f}_\theta - f\|^2 + L^2 |\theta|_q^2 m^{1-2/q}\,.$$

Proof. By Lemma 7.1, for any $h : \mathcal{X} \to \mathbb{R}$ and any $\theta'' \in \mathbb{R}^M$ there exists $\theta' \in \mathbb{R}^M$ such that $|\theta'|_0 \le m$ and

$$\|\mathsf{f}_{\theta'} - h\|^2 \le \|\mathsf{f}_{\theta''} - h\|^2 + \frac{|\theta''|_1^2 L^2}{m}\,. \tag{7.1}$$

Take any $\theta \in \mathbb{R}^M$ and let $J \subseteq \{1, \dots, M\}$ be the set of indices corresponding to the m largest in absolute value components of θ. Take any $f : \mathcal{X} \to \mathbb{R}$ and use (7.1) with $\theta'' = \theta_{J^c}$, $h = f - \mathsf{f}_{\theta_J}$ where $\theta_J = (\theta_j I(j \in J),\ j = 1, \dots, M)$. Then (7.1) takes the form

$$\|\mathsf{f}_{\theta' + \theta_J} - f\|^2 \le \|\mathsf{f}_\theta - f\|^2 + \frac{|\theta_{J^c}|_1^2 L^2}{m}\,. \tag{7.2}$$

Set $\bar{\theta} = \theta' + \theta_J$. By construction, $|\bar{\theta}|_0 \le 2m$. Finally, note that

$$|\theta|_{(j)} \le \frac{|\theta|_q}{j^{1/q}} \tag{7.3}$$

where $|\theta|_{(j)}$ is the jth largest absolute value of the components of θ. Using (7.3) we get the following bound which together with (7.2), yields the lemma:

$$|\theta_{J^c}|_1 = \sum_{j \ge m+1} |\theta|_{(j)} \le |\theta|_{(m)}^{1-q} \sum_{j \ge m+1} |\theta|_{(j)}^q \le \left(\frac{|\theta|_q}{m^{1/q}}\right)^{1-q} |\theta|_q^q = |\theta|_q m^{1-1/q}.$$

□

Lemmas 7.1 and 7.2 combined with Theorem 5.1 imply the following result.

Theorem 7.3. *Assume that $\|f_j\| \le 1$, $j = 1, \dots, M$, and $0 < q \le 1$. Let $\hat f^{ES}$ be the exponential screening estimator with $\beta = 8\sigma^2$ and let $\hat\theta^{ES}$ denote the corresponding vector of weights. Then for any f, and any $\delta > 0$, and integers $n \ge 1$, $M \ge 2$,*

$$E\|\hat f^{ES} - f\|^2 \le \min_{\theta \in B_q(\delta)} \|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(B_q(\delta)) \tag{7.4}$$

where $C > 0$ is an absolute constant, and

$$\psi_{n,M}(B_q(\delta)) = \sigma^{2-q}\delta^q \left[\frac{1}{n}\log\left(1 + \left(\frac{\sigma}{\delta}\right)^q \frac{M}{n^{q/2}}\right)\right]^{1-q/2} \wedge \frac{\sigma^2 R}{n}\,. \tag{7.5}$$

Furthermore, if the model is linear, $y = X\theta + \xi$, then for any $n \ge 1$, $M \ge 2$, $1 \le s \le M$, and $\delta > 0$ we have

$$\sup_{\theta \in B_0(s) \cap B_q(\delta)} \frac{1}{n} E_\theta |X(\hat\theta^{ES} - \theta)|_2^2 \le C\bar\psi_{n,M}(\delta, s, q) \tag{7.6}$$

where $C > 0$ is an absolute constant, and

$$\bar\psi_{n,M}(\delta, s, q) = \psi_{n,M}(B_q(\delta)) \wedge \frac{\sigma^2 s}{n}\log\left(\frac{eM}{s}\right) \wedge \left(\delta^2 + \frac{\sigma^2}{n}\right).$$

Proof. By Theorem 5.1, for an absolute constant $C > 0$ and any $1 \le m \le M/2$,

$$\begin{aligned} E\|\hat f^{ES} - f\|^2 &\le \min_{\theta \in \mathbb{R}^M} \left[\|\mathsf{f}_\theta - f\|^2 + \frac{C\sigma^2}{n}|\theta|_0 \log\left(\frac{eM}{|\theta|_0}\right)\right] + \frac{C\sigma^2}{n} \\ &\le \min_{\theta : |\theta|_0 \le 2m} \|\mathsf{f}_\theta - f\|^2 + \frac{C\sigma^2 m}{n}\log\left(\frac{eM}{2m}\right) \end{aligned}$$

where we have used the monotonicity of the mapping $x \mapsto x\log(eM/x)$. This and Lemma 7.2 imply

$$E\|\hat f^{ES} - f\|^2 \le \min_{\theta \in B_q(\delta)} \|\mathsf{f}_\theta - f\|^2 + C\left(\delta^2 m^{1-2/q} + \frac{\sigma^2 m}{n}\log\left(\frac{eM}{m}\right)\right). \tag{7.7}$$

Minimizing the right hand side of (7.7) in m we obtain the first term on the right hand side of (7.5). The minimum with $\sigma^2 R/n$ comes from Theorem 5.1. Thus (7.4) follows. To show (7.6), we note that replacing the minimum on the right hand side of (5.5) by the value at $\theta = 0$ and using that $f = \mathsf{f}_\theta$ for $\theta \in B_q(\delta)$ yields

$$\sup_{\theta \in B_q(\delta)} \frac{1}{n} E_\theta |X(\hat\theta^{ES} - \theta)|_2^2 \le \sup_{\theta \in B_q(\delta)} \|\mathsf{f}_\theta\|^2 + \frac{C\sigma^2}{n} \le \delta^2 + \frac{C\sigma^2}{n} \tag{7.8}$$

where we have used that $\|\mathsf{f}_\theta\| \le |\theta|_1 \max_j \|f_j\| \le \delta$ for $\theta \in B_q(\delta)$ if $0 \le q \le 1$. Finally, (7.6) is straightforward in view of the last display, (7.4), and (6.2). □

The rates on ℓ_q-balls with $0 < q \le 1$ follow from (7.6) by setting there $s = M$ (then $B_0(s) = \mathbb{R}^M$). Note also that the upper bounds of Theorem 7.3 are optimal in a minimax

sense, cf. Section 8 below. Theorem 7.3 shows that the estimator $\hat{\theta}^{ES}$ attains minimax rates over all $B_0(s) \cap B_q(\delta)$ with $0 < q \leq 1$ (cf. (7.6)), adaptively to s, δ, q, and, in addition, $\hat{\theta}^{ES}$ achieves optimal rates of aggregation on these sets (cf. (7.4)). For $1 < q \leq 2$ we only show that $\hat{\theta}^{ES}$ accomplishes the first task – minimax rates over ℓ_q-balls, under the boundedness assumption on the maximal eigenvalue $\lambda_{\max}(X^T X/n)$ of matrix $X^T X/n$.

Theorem 7.4. *Assume that $\lambda_{\max}(X^T X/n) \leq L^2$, and $1 < q \leq 2$. Let $\hat{\theta}^{ES}$ denote the vector of weights of the exponential screening estimator with $\beta = 8\sigma^2$. If the model is linear, $y = X\theta + \xi$, then for any $n \geq 1$, $M \geq 2$, $1 \leq s \leq M$, and $\delta > 0$ we have*

$$\sup_{\theta \in B_0(s) \cap B_q(\delta)} \frac{1}{n} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2 \leq C \bar{\psi}_{n,M}(L\delta, s, q) \tag{7.9}$$

$C > 0$ is an absolute constant.

Proof. In this case, we get the analog (7.8) with $L\delta$ instead of δ since $\|\mathsf{f}_\theta\|^2 \leq L^2 |\theta|_2^2 \leq (L\delta)^2$ for $\theta \in B_q(\delta)$, $1 < q \leq 2$. To complete the proof, it suffices to show that, for any $\theta \in \mathbb{R}^M$ there exists $\bar{\theta} \in \mathbb{R}^M$ such that $|\bar{\theta}|_0 \leq m$ and

$$|X(\bar{\theta} - \theta)|_2^2/n = \|\mathsf{f}_{\bar{\theta}} - \mathsf{f}_\theta\|^2 \leq L^2 |\theta|_q^2 m^{1-2/q} . \tag{7.10}$$

This replaces Lemma 7.2 when the model is linear, i.e., $f = \mathsf{f}_\theta$. Given (7.10), the argument follows the same lines as in the proof of (7.6). To prove (7.10), take $\bar{\theta} = \theta_J$ where $J \subseteq \{1, \ldots, M\}$ is the set of indices corresponding to the m largest in absolute value components of θ. Then $\|\mathsf{f}_{\bar{\theta}} - \mathsf{f}_\theta\|^2 \leq \lambda_{\max}(X^T X/n) |\bar{\theta} - \theta|_2^2 \leq L^2 |\theta_{J^c}|_2^2$. Using (7.3) we deduce (7.10) from the chain of inequalities

$$|\theta_{J^c}|_2^2 = \sum_{j \geq m+1} |\theta|_{(j)}^2 \leq |\theta|_{(m)}^{2-q} \sum_{j \geq m+1} |\theta|_{(j)}^q \leq \left(\frac{|\theta|_q}{m^{1/q}} \right)^{2-q} |\theta|_q^q = |\theta|_q^2 m^{1-2/q}. \qquad \square$$

8. Minimax lower bounds

The rates for the minimax risk on the intersection of ℓ_0- and ℓ_q-balls obtained in the previous section are optimal if $\delta \geq c_* \sigma/\sqrt{n}$ where $c_* > 0$ is a constant as follows from the next theorem.

Theorem 8.1. *Let $M \geq 1, n \geq 1, 1 \leq s \leq M$, $M \leq n$, and $\delta > 0$. Let either $\Theta_{\delta,s,q} = B_0(s) \cap B_q(\delta)$ and $0 < q \leq 1$ or $\Theta_{\delta,s,q} = \delta \Lambda^M \cap B_0(s)$ and $q = 1$. Then there exists a dictionary $f_1, \ldots, f_M$ with $\max_{1 \leq j \leq M} \|f_j\| \leq 1$ such that*

$$\inf_{\hat{f}} \sup_{\theta \in \Theta_{\delta,s,q}} E_\theta \|\hat{f} - \mathsf{f}_\theta\|^2 \geq C \tilde{\psi}_{n,M}(\delta, s, q)$$

where $C > 0$ is a constant independent of n, M, δ, s, and

$$\tilde{\psi}_{n,M}(\delta, s, q) = \psi_{n,M}(B_q(\delta)) \wedge \frac{\sigma^2 s}{n} \log\left(\frac{eM}{s} \right) \wedge \delta^2 .$$

The proof of Theorem 8.1 is given in [31]($q = 1$), and in [38]($0 < q < 1$). These papers also describe the additional conditions on the matrix X, for which the lower bound holds when not necessarily $M \le n$. Some bounds on the sparse eigenvalues of $X^T X/n$ are then required. Minimax lower bounds for the case $\delta = \infty$, corresponding to the class $B_0(s)$ are studied in [2, 6, 28, 36]. The other extreme case $s = M$, corresponding to the class $B_q(\delta)$, $0 < q \le 1$, is studied in [28] under specific asymptotics on n, M, δ that do not provide the general form of $\tilde{\psi}_{n,M}(\delta, M, q)$; related results are given in [40] for different risk. For the diagonal case when $X^T X/n$ is the identity matrix and $M = n$, upper and lower bounds under some specific asymptotics separately on $B_0(s)$ and on $B_q(\delta)$ are proved in [1, 16, 17]; they are extended to non-asymptotic bounds in [4, 21].

Remark 8.2. If we assume that $\max_{1\le j\le M} \|f_j\| \le L$ instead of $\max_{1\le j\le M} \|f_j\| \le 1$, the lower bound of Theorem 8.1 remains valid with δ replaced by $L\delta$. This remark concerns also the upper bounds of Theorem 7.3.

9. Nonparametric estimation and group sparsity

Consider now Setting 2 of the Introduction (nonparametric regression). Assume that $f \in \mathcal{F}_{\beta,L}$ and the class $\mathcal{F}_{\beta,L}$ is such that assumption (1.7) holds. Theorem 5.1 with $M = n$, and this assumption imply that

$$E\|\hat{f}^{ES} - f\|^2 \le C(\beta, L)^2 k^{-2\beta} + \frac{C\sigma^2}{n} k \log n \tag{9.1}$$

for any $k \le n$. Minimizing this bound in k and taking the suprema we obtain

$$\sup_{f\in\mathcal{F}_{\beta,L}} E\|\hat{f}^{ES} - f\|^2 \le C'(\beta, L) \left(\frac{\log n}{n}\right)^{-2\beta/(2\beta+1)} \tag{9.2}$$

where $C'(\beta, L)$ is a constant depending only on β and L. Thus, the estimator $\hat{f}^{ES}$ attains (adaptively in β, L) the rate $n^{-2\beta/(2\beta+1)}$ up to a logarithmic factor. Note that $n^{-2\beta/(2\beta+1)}$ is the optimal rate of convergence of the squared risk for major classes $\mathcal{F}_{\beta,L}$ satisfying assumption (1.7) [34]. The extra logarithmic factor in (9.2) can be avoided by using, instead of $\hat{f}^{ES}$, a group exponential weighted aggregate, which is of independent interest and is defined as follows.

Let $B_1, \dots, B_K$ be given subsets of $\{1, \dots, M\}$ called the groups. Consider $\theta \in \mathbb{R}^M$ such that $\mathrm{supp}(\theta) \subseteq B = \bigcup_{k=1}^K B_k$ where $\mathrm{supp}(\theta)$ is the set of indices of non-zero components of θ. For any such θ, we denote by $J(\theta)$ a subset of $\{1, \dots, K\}$ of smallest cardinality among all J such that $\mathrm{supp}(\theta) \subseteq B_J = \bigcup_{k\in J} B_k$. Define

$$g(\theta) = |J(\theta)|, \quad B(\theta) = \Big| \bigcup_{k\in J(\theta)} B_k \Big|$$

where $|\cdot|$ denotes the cardinality. For any subset J of $\{1, \dots, K\}$, denote by p^J the sparsity pattern in $\mathcal{P}$ with coordinates

$$p_j^J = \begin{cases} 1 & \text{if } j \in \bigcup_{k\in J} B_k, \\ 0 & \text{otherwise,} \end{cases}$$

for $j = 1, \dots, M$. Consider the set of all such sparsity patterns:

$$\mathcal{P}_{\text{G}} = \left\{ p^J \, , J \subseteq \{1, \dots, K\} \right\} \, .$$

To each sparsity pattern $p \in \mathcal{P}_{\text{G}}$, we assign a least squares estimator $\hat{\theta}_p$ constructed from the first subsample $y^{(1)}$, cf. (5.1). Define the following prior probability distribution on $\mathcal{P}_{\text{G}}$:

$$\pi_p^{\text{G}} = \left[\binom{K}{|J|} e^{|J|} H' \right]^{-1} \, , \quad \forall \, p = p^J, \quad J \subseteq \{1, \dots, K\},$$

where $H' = \sum_{k=0}^{K} e^{-k}$, and consider the exponentially weighted aggregate

$$\hat{f}^{\text{G}} = \sum_{p \in \mathcal{P}_{\text{G}}} \hat{\theta}_p^{\text{G}} \mathsf{f}_{\hat{\theta}_p} \, , \tag{9.3}$$

where $\hat{\theta}^{\text{G}} = (\hat{\theta}_p^{\text{G}}, \ p \in \mathcal{P}_{\text{G}})$ is a vector with components

$$\hat{\theta}_p^{\text{G}} = \frac{\exp(-n\hat{r}_p/\beta)\pi_p^{\text{G}}}{\sum_{p' \in \mathcal{P}} \exp(-n\hat{r}_{p'}/\beta)\pi_{p'}^{\text{G}}}, \quad \forall p \in \mathcal{P}_{\text{G}}.$$

Theorem 9.1. [32] *Let $\beta = 8\sigma^2$. Then for any f, $f_1, \dots, f_M$, and any integers $n \geq 1$, $M \geq 1$, we have*

$$E\|\hat{f}^{\text{G}} - f\|^2 \leq \inf_{\substack{\theta \in \mathbb{R}^M : \\ \text{supp}(\theta) \subseteq B}} \left\{ \|\mathsf{f}_\theta - f\|^2 + \frac{C\sigma^2}{n} \left(B(\theta) + g(\theta) \log \left(\frac{eK}{g(\theta)} \right) + 1 \right) \right\} \tag{9.4}$$

where $C > 0$ is an absolute constant.

Remark that Theorem 9.1 is stated for arbitrary groups B_j. They can overlap and not necessarily cover the whole set $\{1, \dots, M\}$.

Using (9.4), one can prove that the aggregate $\tilde{f}^{\text{G}}$ achieves the optimal rate of convergence under the group sparsity setting [32]. Note that upper bounds for the risk of the Group Lasso estimators in [20, 26] as well as in the earlier papers cited therein depart from this optimal rate at least by a logarithmic factor. Moreover, they are obtained under strong assumptions on the dictionary $\{f_1, \dots, f_M\}$ such as restricted isometry or restricted eigenvalue type conditions, while (9.4) is valid under no assumption on the dictionary.

We now apply Theorem 9.1 for Setting 2 of the Introduction. Let $M = n$, and let all groups B_j be of the same size $T = \lceil (\log n)^2 \rceil$ and form a partition of $\{1, \dots, n\}$, so that, w.l.o.g., $n = KT$. Let the class $\mathcal{F}_{\beta,L}$ be such that (1.7) holds. Denote by $\tilde{f}^{\text{G}}$ the estimator (9.3) with this choice of parameters. In particular, functions f_j are those from (1.7). Fix any $f \in \mathcal{F}_{\beta,L}$. Set $k_* = \lceil n^{1/(2\beta+1)} \rceil$, and let θ^* be the vector in $\mathbb{R}^n$ whose first k_* components are the values $\theta_1^*, \dots, \theta_{k_*}^*$ from (1.7), and other components are 0. Then $g(\theta^*) \leq k_*/T + 1$, and $B(\theta^*) \leq k_* + T$. Plugging these values into the right hand side of (9.4) and using (1.7) with $k = k_*$, we find

$$E\|\tilde{f}^{\text{G}} - f\|^2 \leq C(\beta, L)^2 k_*^{-2\beta} + \frac{C\sigma^2(k_* + T)}{n} \left(1 + \frac{1}{T} \log \left(\frac{en}{k_* + T} \right) \right), \tag{9.5}$$

which immediately implies the next corollary.

Corollary 9.2. *For any class of functions $\mathcal{F}_{\beta,L}$ such that* (1.7) *holds, we have*

$$\sup_{f\in\mathcal{F}_{\beta,L}} E\|\tilde{f}^{G} - f\|^2 \le c(\beta, L, \sigma^2) n^{-2\beta/(2\beta+1)} \tag{9.6}$$

where $c(\beta, L, \sigma^2)$ is a constant depending only on β, L, and σ^2.

Remark that (9.2) and (9.6) are adaptive results. Indeed, the estimators $\hat{f}^{ES}$ and $\tilde{f}^{G}$ do not depend on the parameters β and L, and satisfy these upper bounds simultaneously for all classes $\mathcal{F}_{\beta,L}$ such that (1.7) holds.

10. Universal aggregation

Along with the three main types of aggregation (MS, C, L) described in the Introduction, two other natural examples are of interest: the s-sparse aggregation (L_s) [6], and the convex s-sparse aggregation (C_s) [25]. As summarized in Table 10.1, the sets Θ for the five types of aggregation are either ℓ_0-balls or intersections of ℓ_0-balls with the simplex Λ^M.

Problem	Θ	Description of the oracle
(MS)	$\Theta_{(\mathrm{MS})} = B_0(1) \cap \Lambda^M$	Best in dictionary
(C)	$\Theta_{(\mathrm{C})} = \Lambda^M$	Best convex combination
(L)	$\Theta_{(\mathrm{L})} = \mathbb{R}^M = B_0(M)$	Best linear combination
(L_s)	$\Theta_{(\mathrm{L}_s)} = B_0(s)$	Best s-sparse linear combination
(C_s)	$\Theta_{(\mathrm{C}_s)} = B_0(s) \cap \Lambda^M$	Best s-sparse convex combination

Table 10.1.

The next theorem follows from (6.1), (7.4) with $q = \delta = 1$ and the inclusion $\Lambda^M \subset B_1(1)$.

Theorem 10.1. [31] *Assume that $\max_{1\le j\le M} \|f_j\| \le 1$. Then, for any f, any $M \ge 2$, $n \ge 1$, $1 \le s \le M$, and $\Theta \in \{\Theta_{(\mathrm{MS})}, \Theta_{(\mathrm{C})}, \Theta_{(\mathrm{L})}, \Theta_{(\mathrm{L}_s)}, \Theta_{(\mathrm{C}_s)}\}$ the exponential screening estimator with $\beta = 8\sigma^2$ satisfies the following oracle inequality*

$$E\|\hat{f}^{ES} - f\|^2 \le \min_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta)\,.$$

Here, $C > 0$ is an absolute constant,

$$\psi_{n,M}(\Theta) = \psi^*_{n,M}(\Theta) \wedge \frac{\sigma^2 R}{n}$$

and $\psi^*_{n,M}(\Theta)$ is given in Table 10.2.

Theorem 8.1 implies that the rates $\psi_{n,M}(\Theta)$ given in Theorem 10.1 are optimal rates of aggregation for the corresponding classes Θ. Indeed, to show (1.12) with the same rates, it suffices to use the lower bounds for the minimax risk, since obviously

$$\inf_{\hat{f}} \sup_{f} \left(E\|\hat{f} - f\|^2 - \min_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 \right) \ge \inf_{\hat{f}} \sup_{\theta\in\Theta} E_\theta \|\hat{f} - \mathsf{f}_\theta\|^2. \tag{10.1}$$

On the other hand, the sets Θ in Theorem 10.1 are either ℓ_0-balls or intersections of ℓ_0-balls with the ℓ_1-simplex Λ^M, and the lower bounds for the minimax risk on these sets are available from Theorem 8.1.

Problem	$\psi^*_{n,M}(\Theta)$
(MS)	$\frac{\sigma^2 \log M}{n}$
(C)	$\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{\sqrt{n}}\right)}$
(L)	$\frac{\sigma^2 R}{n}$
(L_s)	$\frac{\sigma^2 s}{n} \log\left(\frac{eM}{s}\right)$
(C_s)	$\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{\sqrt{n}}\right)} \wedge \frac{\sigma^2 s}{n} \log\left(\frac{eM}{s}\right)$

Table 10.2.

If assumption $\max_{1\le j\le M} \|f_j\| \le 1$ in Theorem 10.1 is replaced by $\max_{1\le j\le M} \|f_j\| \le L$, the rates in Table 10.2 remain valid with the only difference that $\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{\sqrt{n}}\right)}$ should be replaced by $L\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{L\sqrt{n}}\right)}$.

We see that the problem of aggregation is closely related to that of minimax estimation on the intersection of ℓ_0- and ℓ_1-balls. Indeed, for both problems upper bounds for the risk and for the excess risk are attained by one and the same estimator, which is the exponential screening estimator. Furthermore, the optimal rates of aggregation in Theorem 10.1 are similar to the minimax rates on the intersection of the corresponding ℓ_0- and ℓ_1-balls (cf. Section 7).

Using (10.1), the upper bounds (6.1), (7.4), and Theorem 8.1 we also find that the estimator $\hat{f}^{ES}$ attains the optimal rates of ℓ_q-aggregation:

Theorem 10.2. *Let the assumptions of Theorems 8.1 and 10.1 be satisfied, and* $0 < q \le 1$. *Then the estimator* $\hat{f}^{ES}$ *with* $\beta = 8\sigma^2$ *is an optimal aggregate for the classes* $\Theta = B_q(\delta)$, $\delta > 0$. *The optimal rates of aggregation for these classes are* $\psi_{n,M}(B_q(\delta))$. *In addition, this estimator is an optimal aggregate for the classes* $\Theta = B_q(\delta) \cap B_0(s)$ *with* $\delta \ge c_*\sigma/\sqrt{n}$ *where* $c_* > 0$ *is a constant independent of* n, M. *The optimal rates of aggregation for these classes are* $\tilde{\psi}_{n,M}(\delta, s, q)$.

In summary, the exponential screening estimator enjoys the property of *universal aggregation*, i.e., it attains optimal rates of aggregation simultaneously on all the classes Θ considered in this section.

Acknowledgements. This work is supported by GENES, and by the French National Research Agency (ANR) under the grants Idex ANR -11- IDEX-0003-02, Labex ECODEC (ANR - 11-LABEX-0047), and IPANEMA (ANR-13-BSH1-0004-02). The author is grateful to Cowles Foundation and to the Department of Statistics of Yale University for their

hospitality during the writing of this paper.

References

[1] Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M., *Adapting to Unknown Sparsity by Controlling the False Discovery Rate*, Ann. Statist. **34** (2006), 584–653.

[2] Abramovich, F. and Grinshtein, V., *MAP Model Selection in Gaussian Regression*, Electronic J. of Statistics **4** (2010), 932–949.

[3] Alquier, P., Lounici, K., *PAC-Bayesian bounds for sparse regression estimation with exponential weights*, Electronic J. of Statistics **5** (2011), 127–145.

[4] Birgé, L. and Massart, P., *Gaussian model selection*, J. Eur. Math. Soc. **3** (2001), 203–268.

[5] ———, *Aggregation for Regression Learning*, 2004. arXiv:math.ST/0410214.

[6] Bunea, F., Tsybakov, A. B., and Wegkamp, M., *Aggregation for Gaussian Regression*, Annals of Statistics **35** (2007), 1674–1697.

[7] Catoni, O., *Statistical Learning Theory and Stochastic Optimization*, Saint-Flour Summer School in Probability XXXI, 2001. Lecture Notes in Mathematics 1851. Springer, Berlin, 2004.

[8] Cesa-Bianchi, N. and Lugosi, G., *Prediction, Learning, and Games*, Cambridge Univ. Press, 2006.

[9] Dai, D, Rigollet, P., and Zhang, T., *Deviation Optimal Learning using Greedy Q-aggregation*, Annals of Statistics **40** (2012), 1878-1905.

[10] Dai, D, Rigollet, P., Xia, L., and Zhang, T., *Aggregation of Affine Estimators*, Electronic J. of Statistics **8** (2014), 302–327.

[11] Dalalyan, A. S. and Salmon, J., *Sharp Oracle Inequalities for Aggregation of Affine Estimators*, Annals of Statistics **40** (2012), 2327-2355.

[12] Dalalyan, A. and Tsybakov, A. B., *Aggregation by Exponential Weighting and Sharp Oracle Inequalities*, In Proc. COLT 2007, Lecture Notes in Artificial Intelligence 4539, 97–111. Springer, Berlin-Heidelberg, 2007.

[13] ———, *Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity*, Machine Learning **72** (2008), 39–61.

[14] ———, *Mirror Averaging with Sparsity Priors*, Bernoulli **18** (2012), 914–944.

[15] ———, *Sparse Regression Learning by Aggregation and Langevin Monte-Carlo*, Journal of Computer and System Sciences **78** (2012), 1423-1443.

[16] Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S., *Maximum Entropy and the Nearly Black Object*, J. Roy. Statist. Soc. Ser. B **54** (1992), 41–81.

[17] Donoho, D. L. and Johnstone, I. M., *Minimax Risk Over l_p-balls for l_q-error*, Probab. Theory Related Fields **99** (1994), 277–303.

[18] Gerchinovitz, S., *Prediction of Individual Sequences and Prediction in the Statistical Framework : some Links around Sparse Regression and Aggregation Techniques*, Ph.D thesis, Université Paris Sud, 2011.

[19] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A., *Wavelets, Approximation, and Statistical Applications*, Lecture Notes in Statistics 129. Springer, NY, 1998.

[20] Huang, J. and Zhang, T., *The benefit of group sparsity*, Annals of Statistics **38** (2010), 1978–2004.

[21] Johnstone, I., *Gaussian Estimation: Sequence and Wavelet Models*, Draft of a book. http://www-stat.stanford.edu/~imj/GE06-11-13.pdf, 2011.

[22] Koltchinskii, V., Lounici, K., and Tsybakov, A. B., *Nuclear Norm Penalization and Optimal Rates for Noisy Low Rank Matrix Completion*, Annals of Statistics **39** (2011), 2302–2329.

[23] Lecué, G. and Rigollet, P., *Optimal Learning with Q-aggregation*, Annals of Statistics **42** (2014), 211–224.

[24] Leung, G. and Barron, A. R., *Information Theory and Mixing Least-squares Regressions*, IEEE Trans. Inform. Theory **52** (2006), 3396–3410.

[25] Lounici, K., *Generalized Mirror Averaging and D-convex Aggregation*, Mathematical Methods of Statistics **16** (2007), 246–259.

[26] Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S., *Oracle inequalities and optimal inference under group sparsity*, Ann. Statist. **39** (2011), 2164–2204.

[27] Nemirovski, A., *Topics in Non-parametric Statistics*, Saint-Flour Summer School in Probability XXVIII, 1998. Lecture Notes in Mathematics 1738. Springer, NY, 2000.

[28] Raskutti, G., Wainwright, M. J., and Yu, B., *Minimax Rates of Estimation for High-dimensional Linear Regression over l_q-balls*, IEEE Trans. Inform. Th. **57** (2011), 6976–6994.

[29] Rigollet, P., *Kullback-Leibler Aggregation and Misspecified Generalized Linear Models*, Annals of Statistics **40** (2012), 639–665.

[30] Rigollet, P. and Tsybakov, A. B., *Linear and Convex Aggregation of Density Estimators*, Mathematical Methods of Statistics **16** (2007), 260–280.

[31] ———, *Exponential Screening and Optimal Rates of Sparse Estimation*, Annals of Statistics **39** (2011), 731–771.

[32] ———, *Sparse Estimation by Exponential Weighting*, Statistical Science **27** (2012), 558–575.

[33] Tsybakov, A. B., *Optimal Rates of Aggregation*, In Proc. COLT 2003, Lecture Notes in Computer Science 2777, 303–313. Springer, NY, 2003.

[34] _____, *Introduction to Nonparametric Estimation*, Springer, NY, 2009.

[35] _____, *Aggregation and High-dimensional Statistics*, Lecture notes of Saint-Flour Summer School in Probability, July 2013.

[36] Verzelen, N., *Minimax Risks for Sparse Regressions: Ultra-high Dimensional Phenomenons*, Electron. J. Stat. **6** (2012), 38–90.

[37] Vovk, V., *Aggregating Strategies*, In Proc. 3rd Annual Workshop on Computational Learning Theory, 372–383. Morgan Kaufmann, San Mateo, CA, 1990.

[38] Wang, Z., Paterlini, S., Gao, F., and Yang, Y., *Adaptive Minimax Estimation over Sparse l_q-hulls*, 2011. arXiv:1108.1961.

[39] Yang, Y., *Aggregating Regression Procedures to Improve Performance*, Bernoulli **10** (2004), 25–47.

[40] Ye, F. and Zhang, C.-H., *Rate Minimaxity of the Lasso and Dantzig Selector for the ℓ_q loss in ℓ_r balls*, Journal of Machine Learning Research **11** (2010), 3519–3540.

Laboratoire de Statistique, CREST-ENSAE, 3, av. Pierre Larousse 92245 Malakoff, France
E-mail: alexandre.tsybakov@ensae.fr

Operator limits of random matrices

Bálint Virág

Abstract. We present a brief introduction to the theory of operator limits of random matrices to non-experts. Several open problems and conjectures are given. Connections to statistics, integrable systems, orthogonal polynomials, and more, are discussed.

Mathematics Subject Classification (2010). Primary 60B11; Secondary 62H12.

Keywords. Random matrix, random operator.

1. Introduction

Wigner introduced random matrices to mathematical physics as a model for eigenvalues in a disordered system, such as a large nucleus. In the classical approach to random matrices, one considers some statistic of the matrix, and tries to understand the large n limit.

Here we follow a different approach. It is along the lines of the "objective method" coined by David Aldous. The goal is to take a limit of the entire object of interest, in this case the matrix itself. This has the advantage that the structure in the matrix will be preserved in the random limit. This method has been very successful in understanding random objects, notable examples are (the classical) Brownian motion, the continuum random tree, the Brownian map, and SLE, and recent limits of dense and sparse graphs.

This study of random matrices was initiated by the predictions in the work of Edelman and Sutton [19]. They suggested that the tridiagonal matrix models introduced by Trotter [43] and Dumtiriu and Edelman [17], should have certain differential operator limits. Their work was the starting point of intense activity in the area, which is what this paper intends to review.

We will first introduce the tridiagonal models. Then we consider various operator limits and discuss some applications.

2. Tridiagonal models

Trotter never thought that his 1984 paper [43], in which he introduced tridiagonalization to the theory of random matrices, would ever be very important. Indeed, he just used it to give a different proof for the Wigner semicircle law for the GOE, of which there are (and had been) a plethora of other proofs. His proof was nevertheless beautiful, and we will present a quick modern version in Section 3.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Tridiagonalization is a method to find eigenvalues of self-adjoint matrices that is still used in modern software, for example in the Lanczos algorithm. It is also useful if we want to store the eigenvalues of an $n \times n$ matrix, but not n^2 data points, without operations beyond linear algebra.

Starting with an $n \times n$ symmetric matrix A, first conjugate it with a special block orthogonal matrix so that its first coordinate vector is fixed. Writing both matrices in the block form

$$\begin{pmatrix} 1 & \\ & O \end{pmatrix} \begin{pmatrix} a & b^\dagger \\ b & C \end{pmatrix} \begin{pmatrix} 1 & \\ & O^\dagger \end{pmatrix} = \begin{pmatrix} a & (Ob)^\dagger \\ Ob & OCO^\dagger \end{pmatrix}$$

so one can choose O so that b becomes a nonnegative multiple of the first coordinate vector, and the first row is like that of a tridiagonal matrix. One can iterate this procedure (conjugating by an orthogonal matrix fixing the first k coordinates in the kth step), to get a tridiagonal matrix.

The Gaussian orthogonal ensemble (GOE) is the random matrix $A = (M + M^t)/\sqrt{2}$ where M has independent standard Gaussian entries. It has the property that conjugation by an orthogonal matrix preserves its distribution.

Exploiting this property and independence, we see that the result of tridiagonalization is a symmetric matrix with independent diagonals a_i, (resp. off-diagonals b_i). Setting $\beta = 1$ and dividing by $\sqrt{n\beta}$ we get the tridiagonal matrix T with entries

$$a_i \sim N(0, 2/n\beta), \qquad b_i \sim \chi_{(n-i)\beta}/\sqrt{n\beta}. \tag{2.1}$$

(Recall that χ_k is the distribution of the length of an n-dimensional vector with independent standard normal entries). Starting with standard complex normals gives the Gaussian unitary ensemble (GUE) and the same story with $\beta = 2$. It will be convenient to consider the resulting joint density for the variables a_i, $\log b_i$ as a constant times

$$\exp(-\tfrac{\beta}{4} n \operatorname{tr} V(T)) \times \prod_{k=1}^{n-1} b_k^{\beta(n-k)} \tag{2.2}$$

with $V = x^2$.

The tridiagonalization procedure seem to produce a non-unique result (there are many choices for the orthogonal matrices), but this is not the case. If the vectors $e, Ae, \dots A^{n-1}e$ are linearly independent, we always get the *same* Jacobi matrix (tridiagonal with positive off-diagonals). It is, in fact the matrix A written in the Gram-Schmidt orthonormalization of this basis.

In both descriptions, T is an orthogonal conjugate to A, with the first coordinate vector fixed. If one defines this as an equivalence relation on symmetric matrices where e is cyclic, then each class contains exactly one Jacobi matrix, so they are natural class representatives.

So T, with $2n-1$ parameters, encodes the n eigenvalues of A. But what else does this encode? Check that

$$A_{11}^k = T_{11}^k = \int x^k d\sigma,$$

for the **spectral measure**

$$\sigma = \sum_{i=1}^{n} q_i \delta_{\lambda_i},$$

where $q_i = \varphi_{i,1}^2$ for the normalized eigenvectors φ_i. So T encodes the spectral measure, which is a probability measure supported on n points and so are described by $2n-1$ parameters.

Since for the GOE the eigenvectors are uniform on the unit n-sphere and independent of the eigenvalues, we can write the joint density on $\lambda_i, \log q_i$ as a constant times

$$\exp(-\beta n \operatorname{tr} V(A)) \times \prod_{i<j} |\lambda_i - \lambda_j|^\beta \prod_{k=1}^n q_i^{\beta/2} \tag{2.3}$$

using the well-known formula for the eigenvalue distribution [1]. Now the factors the left of $\times$ in (2.2) and (2.3) are equal, since A, T have the same eigenvalues. Interestingly, the same holds for the value on the right, see Section 3.1 of [12]! Since it is also known that the map

$$(a_1, \ldots, a_n, \log b_1, \ldots \log b_{n-1}) \mapsto (\lambda_1, \ldots, \lambda_n, \log q_1, \ldots \log q_n) \tag{2.4}$$

is a bijection, it follows that it is **measure-preserving** (up to a fixed constant). As a consequence, the equivalence of measures (2.2), (2.3) holds for all functions V and $\beta > 0$. When $V = x^2$, the model is called the β-Hermite ensemble and this was shown with the same methods by Dumitriu and Edelman [17]. Just as in the special cases of the GOE and GUE, the tridiagonal matrix T has independent entries.

This model (2.3) on n points is called Dyson's beta ensemble.

Structure of the tridiagonal matrices. As one expects, various features of the eigenvalue distribution can be read off the tridiagonal matrix T. For example, the top (and bottom) eigenvectors of the matrix have all of their ℓ^2 mass in the first order $n^{1/3}$ coordinates. So in order to understand edge statistics, one can take a scaling limit of this part of T.

Similarly, for the β-Hermite T eigenvectors for eigenvalues near 0 have their ℓ^2-mass distributed through the whole length n. So bulk local statistics of eigenvalues will be understood by taking an operator limit of T on this scale.

So while local eigenvalue statistics have to do with the global structure of T, the global statistics of eigenvalues (such as the Wigner semicircle law) have to do with the local structure of T at a random vertex, as we will see next. The spectral measure at the first coordinate is also closely related to the eigenvalue distribution.

3. Density of states

In this section, we pursue the point of view of operator limits to deduce the Wigner semicircle law. In fact, we will get two proofs, one using rooted convergence of graphs, and the other using Benjamini-Schramm convergence.

Rooted convergence and the Wigner semicircle law. A sequence of edge-labeled, bounded degree rooted graphs (G_n, o) is said to converge locally to a rooted graph G if for every r, the r neighborhood of o the graph stabilizes and the labels in the neighborhood converge pointwise as $n \to \infty$.

For example, using the asymptotics

$$\chi_n \approx \sqrt{n} + N(0, 1/2),$$

we see that the β-Hermite ensemble matrix $T = T_n$ (thought of as weighted adjacency matrix) rooted at the first vertex converges almost surely locally to the graph T^* of the nonnegative integers (with weights 1) as $n \to \infty$ and β is fixed.

Here we identity the graphs with their adjacency matrices. Recall the spectral measure of G at o is the measure whose k-th moments are $G^k_{o,o}$. The method of moments shows that rooted convergence implies convergence of spectral measures at the root o.

The moments of the spectral measure of T^* at o these are the number of returning simple random walk paths that stay nonnegative; they characterize the Wigner semicircle law.

What we have shown is that the spectral measures converge almost surely. But the spectral measure assigns Dirichlet$(\beta/2, \dots \beta/2)$ weights to the eigenvalues, see (2.3). The law of large numbers for these weights shows that the empirical eigenvalue distribution has the same limit.

An argument like this works for more general potentials V – in this case the limiting rooted labeled graph is the Jacobi operator associated to the orthogonal polynomials with respect to the measure $e^{-V(x)}\,dx$, see [31].

Benjamini-Schramm limits and the Wigner semicircle law. Here we deduce the semicircle law in a way which is, essentially, equivalent to Trotter's [43] but uses no computation. A sequence of unrooted, labeled finite graphs G_n is said to converge to a random rooted graph (G, o) in the Benjamini-Schramm sense if the law of (G_n, o) converges there with uniform choice of o. The convergence is with respect to the topology of rooted convergence introduced above.

Again, the method of moments shows that the expected spectral measure at o, which is the empirical eigenvalue distribution of G_n, converges to the expected spectral measure of (G, o) at o.

A moment of thought shows that the almost sure Benjamini-Schramm limit of the β-Hermite ensembles is $\sqrt{U}\mathbb{Z}$, where $\mathbb{Z}$ is the graph of the integers, rooted at o, U is a uniform random variable that comes from the mean of the χ variable at the uniformly chosen location of the root.

Now $\mathbb{Z}$ is also the Benjamini-Schramm limit of n-cycles, whose eigenvalues are the real parts of equally spaced points on the circle $\{|z| = 2\} \subset \mathbb{C}$. Hence the spectral measure of $\mathbb{Z}$ is the real part of uniformly chosen point on the circle of radius 2.

The expected spectral measure μ of $\sqrt{U}\mathbb{Z}$ is thus the real part of the uniformly chosen point from a random circle with radius $2\sqrt{U}$; but this is just another way to chose a point from uniform measure in the disk of radius two. Thus μ is the semicircle law.

4. The β-Hermite random measure on $\mathbb{R}$

A special property of the β-Hermite matrices $\sqrt{n}\,T_n$ is that they are minors of each other; as a result, they are the minor of a semi-infinite Jacobi matrix $J = J_\beta$.

The $\beta \to \infty$ limit J_∞ has zeros on the diagonal and $\sqrt{k}$ at positions $(k+1, k)$ and $(k, k+1)$. Its spectral measure at the first coordinate is standard normal.

Such matrices have relevance in the theory of orthogonal polynomials. Here we review a few brief facts. Given a measure μ with infinite support on $\mathbb{R}$ with sufficiently thin tails, the kth orthonormal polynomial is the unique degree n polynomial with positive main coefficient that is orthogonal in $L^2(\mathbb{R}, \mu)$ to all lower degree polynomials.

One can show that there are unique a_n and $b_n > 0$ so that the p_n satisfy a recursion $p_{k-1}b_{k-1} + p_k a_k + p_{k+1}b_k = xp_k$. In other words, the (not necessarily ℓ^2) vector $p(x) = (p_k(x))_{k\geq 0}$ satisfies the eigenvector equation $Jp(x) = xp(x)$ where J is the infinite tridiagonal matrix built from the a-s and b-s. Note that here it is crucial that the numbering is reversed compared to (2.1).

Note that $p(x)$ restricted to the first n coordinates is an eigenvector of the $n \times n$ minor of J if and only if $p_n(x) = 0$. In particular, the p_n are constant multiples of the characteristic polynomials of this minor.

Conversely, given such J and assuming that it is self-adjoint, one can recover the measure μ as the spectral measure of J at the first coordinate. Since J_β is easily shown to be self-adjoint, we have shown

Theorem 4.1 (Coupling of the β-Hermite ensembles). *There exists a random measure μ_β so that for all n the zeros of the orthogonal polynomial p_n with respect to μ_β are distributed as the eigenvalues of the n-point β-Hermite ensemble.*

It also follows that the β-Hermite eigenvalues are exactly the Gaussian quadrature points for this measure!

The measure μ_β can be thought of as a random "rough" version of the standard normal distribution (μ_∞). The measure has been studied by Breuer, Forrester, and Smilansky [8]. They showed that its Hausdorff dimension is almost surely equal to $(1 - 2/\beta)^+$. For $\beta < 2$, the measure is pure point. A similar phenomenon holds for the family of Gaussian multiplicative cascade measures, see, for example [40] in some sense it is a noncommutative version. A natural question is the following

Question 1 (Spectral measure and multiplicative cascades). *Does the β-Hermite measure and the Gaussian multiplicative cascade measure with the same Hausdorff dimension have the same fractal spectrum?*

Question 2 (Nested models). *Can any other Dyson β-ensembles be coupled this way? How about other natural random matrix models?*

5. Edge limits and the stochastic Airy operator

For n large and $k = o(n)$, we have the asymptotics $\chi_{n-k} \asymp \sqrt{n} - k/\sqrt{4n} + N(0, 1/2)$. Thus the top minor of size $o(n)$ of $(2I - T)$ looks like a discrete second derivative plus multiplication by $2k/n$, plus multiplication by discrete independent noise. The precise continuous analogue would be

$$\mathrm{SAO}_\beta = -\partial_t^2 + t + \tfrac{2}{\sqrt{\beta}} b' \tag{5.1}$$

called the Stochastic Airy Operator, where b' is a distribution (the derivative of standard Brownian motion). Edelman and Sutton [19] conjectured that this operator, acting on $L^2(\mathbb{R}^+)$ with Dirichlet boundary conditions $f(0) = 0$, is the edge limit of T_n. This was proved in in [38]:

Theorem 5.1. *There exists a coupling of the β-Hermite random matrices T_n on the same probability space so that a.s. we have*

$$n^{2/3}(2I - T_n) \to \mathrm{SAO}_\beta$$

in the norm-resolvent sense: for every k the bottom kth eigenvalue converges the and corresponding eigenvector converges in norm. Here $2I-T_n$ acts on the embedding $\mathbb{R}^n \subset L^2(\mathbb{R}_+)$ with coordinate vectors $e_j = n^{1/6}\mathbf{1}_{[j-1,j]n^{-1/3}}$.

The limiting distribution of the top eigenvalue of the GOE, and GUE are called the Tracy-Widom distribution TW_β with $\beta = 1, 2$, respectively. It follows that for $\beta = 1, 2$ the negative of the bottom eigenvalue $-\Lambda_0$ of SAO_β has TW_β distribution. For more general β, this can be taken as a definition of $\mathrm{TW}\beta$.

The domain of SAO_β can be defined precisely (see [4]), but we will not do that here. The eigenvalues and eigenvectors can be defined though the Courant-Fisher characterization,

$$\Lambda_k = \inf_{A:\dim A=k+1} \sup_{f\in A, \|f\|_2=1} \langle f, \mathrm{SAO}_\beta f\rangle.$$

the latter can be defined via integration by parts as long as f, f' and $\sqrt{t}f$ are in $L^2(\mathbb{R}^+)$, and in the formula A is a subspace of such functions. The eigenvectors are defined as the corresponding minimizers, and can be shown to be unique, see [38].

Glimpses of the proof of Theorem 5.1. We explain how to show that the bottom eigenvalue converges (see [38] for the rest). It is a nice exercise [38] to show that given a Brownian path and $\varepsilon > 0$ there is a random constant C so that for every function f with $f, f', \sqrt{t}f \in L^2(\mathbb{R})$ we have

$$\left|\int f^2\, dB\right| \le C\|f\|^2 + \varepsilon(\|f'\|^2 + \|\sqrt{t}f\|^2) = C\|f\|^2 + \varepsilon\langle f, \mathrm{AO}f\rangle.$$

where $\mathrm{AO} = \mathrm{SAO}_\infty$ is the usual Airy operator $-\partial_t^2 + t$. In other words, we have the positive definite order of operators

$$-C + (1-\varepsilon)\mathrm{AO} \le \mathrm{SAO}_\beta \le (1+\varepsilon)\mathrm{AO} + C \tag{5.2}$$

Using Skorokhod's representation and the central limit theorem, we can guarantee a coupling so that the integrated potential of $2I - T_n$ converges uniformly on compacts to that of SAO_β. Moreover, the discrete analogues of the bound (5.2) will hold with uniform constants C and all n. Note that taking the bottom eigenvector f_0 of SAO_β and plugging it into the approximating operators, the Rayleigh quotient formula shows that their bottom eigenvalues satisfy

$$\limsup \lambda(n) \le \Lambda_0$$

Conversely, SAO_β can be tested against any weak limit of the bottom eigenfunctions $f(n)$, which must exist because of the discrete version of (5.2) guarantees enough tightness. As a result,

$$\liminf \lambda(n) \ge \Lambda_0. \qquad \square$$

A different operator appears at the so-called hard edge, see [37, 39] for further analysis.

6. Applications of the stochastic Airy operator

The stochastic Airy operator is a Schrödinger-type operator, and therefore tools from the classical theory are applicable.

First, as a self-adjoint operator, one can use Rayleigh quotients or positive definite ordering to characterize its low-lying eigenvalues. Second, as a Schrödinger operator, one can use oscillation theory for the same. We will briefly show how these methods work.

Theorem 6.1. *Let $\Lambda_k \uparrow$ be the eigenvalues of* SAO_β. *Then almost surely*

$$\lim_{k\to\infty} \frac{\Lambda_k}{k^{2/3}} = \left(\frac{3\pi}{2}\right)^{2/3}$$

Proof. As a consequence of (5.2), that inequality (5.2) also holds when we replace the operators AO, SAO_β by their $k+1$st eigenvalues $\mathcal{A}_k$, Λ_k. By letting $\varepsilon \to 0$ we see that $\Lambda_k/\mathcal{A}_k \to 1$ a.s. Now note that eigenfunctions of AO are translates of the solution Ai of the Airy differential equation

$$(-\partial_t^2 + t)\,\mathrm{Ai} = 0, \qquad \mathrm{Ai}(t) \to 0 \text{ as } t \to \infty \tag{6.1}$$

by some a so that $\mathrm{Ai}(-a) = 0$. The classical asymptotics of the zeros of Ai now imply the claim. □

Applications of the Rayleigh quotient formula. Next, we show an argument from [38] that gives a sharp bound on the sub-Gaussian left tail of the TW_β distribution of $-\Lambda_0$. It only relies on Rayleigh quotients and standard Gaussian tail bounds!

Lemma 6.2.

$$P(\Lambda_0 > a) \le \exp\Big(-\frac{\beta}{24}\,a^3(1+o(1))\Big).$$

Proof. The Raleigh quotient formula implies that

$$\Lambda_0 > a \qquad \Rightarrow \qquad \langle f, \mathrm{SAO}_\beta f\rangle > a$$

for all nice functions f. Note that any fixed f will give a bound, and $\langle f, \mathrm{SAO}_\beta f\rangle$ is just a Gaussian random variable with mean $\|f'\|_2^2 + \|f\sqrt{t}\|_2^2$ and variance $\frac{4}{\beta}\|f\|_4^4$. In the quest for a good f one expects the optimal f to be relatively "flat" and ignore the $\|f'\|_2^2$ term. In the tradition of zero-knowledge proofs, it is legal to hide the resulting variational problem and how to solve it from the reader (see [38] Section 4). Out of the hat comes

$$f(x) = (x\sqrt{a}) \wedge \sqrt{(a-x)^+} \wedge (a-x)^+,$$

where the middle term is dominant, while the others control $\|f'\|_2$. Then

$$a\|f\|_2^2 \sim \frac{a^3}{2}, \qquad \|f'\|_2^2 = O(a), \qquad \|\sqrt{x}f\|_2^2 \sim \frac{a^3}{6}, \qquad \|f\|_4^4 \sim \frac{a^3}{3}.$$

The proof is completed by substitution, with a standard normal N,

$$P(\Lambda_0 > a) \le P\left(\frac{2}{\sqrt{3\beta}}\,a^{3/2}\,N > a^3\left(\frac{1}{2} - \frac{1}{6} + o(1)\right)\right) = \exp\Big(-\frac{\beta}{24}\,a^3(1+o(1))\Big).$$

□

Applications of Sturm-Liouville oscillation theory. Taking the logarithmic derivative $W = f'/f$ (also called Riccati transformation) transforms the eigenvalue equation $\mathrm{SAO}_\beta f = \lambda f$ to a first order non-linear ODE. We write this in the SDE form

$$dW = \tfrac{2}{\sqrt{\beta}}\, db + \big(t - \lambda - W^2\big)\, dt \tag{6.2}$$

this can be thought of as an equation on the circle compactification of $\mathbb{R}$: a solution that explodes to $-\infty$ in finite time should continue from $+\infty$. In this sense, the solution is monotone in λ: increasing λ moves it the "down" direction on the circle.

Let's first restrict the operator to a finite interval $[0, \tau]$ with Dirichlet boundary condition. Then λ is an eigenvalue iff an explosion happens at τ, and increasing λ moves the explosions to the left. On $(0, \tau)$ we thus have

$$\#\{\text{ explosions }\} \;=\; \#\{\text{ eigenvalues } <\; \lambda\,\}\,. \tag{6.3}$$

For the SAO_β this statement remains true with $\tau = \infty$, and as a consequence

$$P(W_\lambda \text{ never explodes}) = P(\lambda < \Lambda_0).$$

Let $P_{t,w}$ denote the law of the solution W of the $\lambda = 0$ version of (6.2) started at time t and location w. Setting

$$F(t, w) = P_{t,w}(W \text{ never explodes }),$$

we see that the translation invariance of (6.2) implies that

$$\lim_{w \uparrow \infty} F(-\lambda, w) = P(\lambda < \Lambda_0).$$

This gives a characterization for the Tracy-Widom distribution TW_β of $-\Lambda_0$. Boundary hitting probabilities of an SDE can always be expressed as solutions of a PDE boundary value problem. Indeed, such functions are martingales and are killed by the generator, see [5]. So F satisfies

$$\partial_t F + \tfrac{2}{\beta}\, \partial_w^2 F + (t - w^2)\, \partial_w F \;=\; 0 \qquad \text{for } t, w \in \mathbb{R}, \tag{6.4}$$

with $F(t, w) \to 1$ as $t, w \to \infty$ together, and $F(t, w) \to 0$ as $w \to -\infty$ with t bounded above.

It is easy to check that the problem has a unique bounded solution, and so it gives a characterization of the Tracy-Widom-β distribution. However, new ideas were needed to connect these equation to the Painlevé systems; before we turn to these, we consider an application of (6.2) from [38].

SDE representation and tail bounds. We now show how the SDE representation (6.2) is used to attain tail bounds for the law $\mathrm{TW}_\beta = -\Lambda_0$ in [38]. We prove the matching lower bound to Lemma 6.2; readers not familiar with Cameron-Martin-Girsanov transformations may skip this proof.

Lemma 6.3.

$$P(\Lambda_0 > a) \geq \exp\Big(-\frac{\beta}{24}\, a^3 (1 + o(1))\Big).$$

Proof. By monotonicity of the solutions, we have

$$P_{\infty,-a}\Big(W \text{ never explodes}\Big) \geq P_{1,-a}\Big(W \text{ never explodes}\Big)$$
$$\geq P_{0,-a}\Big(W_t \in [0,2] \text{ for all } t \in [-a,0]\Big) P_{0,0}\Big(W \text{ never explodes}\Big).$$

The last factor in line two is some positive number not depending on a. To bound the first factor from below, we first write it using Cameron-Martin-Girsanov formula as

$$E_{1,-a}\Big[\exp\left(-\frac{\beta}{4}\int_{-a}^{0}(t-b_t^2)db_t - \frac{\beta}{8}\int_{-a}^{0}(t-b_t^2)^2dt\right);\, b_t \in [0,2] \text{ for all } t \leq 0\Big],$$

where, for this proof only, b_t denotes a Brownian motion with diffusion coefficient $2/\sqrt{\beta}$. On the event above, the main contribution comes from

$$\frac{\beta}{8}\int_{-a}^{0}(t-b_t^2)^2\,dt = \frac{\beta}{24}a^3 + O(a^2),$$

of lower order is the second term

$$\int_{-a}^{0}(t-b_t^2)db_t = ab_{-a} + \frac{1}{3}(b_{-a}^3 - b_0^3) + (\frac{4}{\beta}-1)\int_{-a}^{0}b_t dt = O(a).$$

We are left to compute the probability of the event

$$P_{-a,0}(b_t \in [0,2] \text{ for } t \leq 0) \geq e^{-ca},$$

since it is the chance of a Markov chain staying in a bounded set for time proportional to a. This does not interfere with the main term. □

In [16] arguments of this kind are used to provide a more precise bound for the other tail $P(\Lambda_0 < -a)$, including $-3/4$ the exponent in the polynomial correction. It was shown that

$$P\left(TW_\beta > a\right) = a^{-\frac{3}{4}\beta+o(1)}\exp\left(-\frac{2}{3}\beta a^{3/2}\right).$$

See [6] for further non-rigorous results in this direction.

Tail estimates for finite n. It is possible to make versions the tail estimate proofs for finite n, before taking the limit. This was carried out by Ledoux and Rider [33]. They give strong tail estimates for the β-Hermite (and also Laguerre) ensembles for finite n. We quote the β-Hermite results from that paper.

Theorem 6.4. *There are absolute constants c, C so that for all $n \geq 1$, $\varepsilon \in (0,1]$ and $\beta \geq 1$ the β-Hermite ensemble T_n satisfies*

$$c^\beta e^{-\beta n\varepsilon^{3/2}/c} \;\leq\; P\Big(\lambda_1(T_n) \geq 2(1+\varepsilon)\Big) \;\leq\; Ce^{-\beta n\varepsilon^{3/2}/C}$$

and

$$c^\beta e^{-\beta n^2\varepsilon^{3}/c} \;\leq\; P\Big(\lambda_1(T_n) \leq 2(1-\varepsilon)\Big) \;\leq\; C^\beta e^{-\beta n^2\varepsilon^{3}/C}$$

For the second lower bound we need to assume in addition that $\varepsilon < c$.

7. Finite rank perturbations and Painlevé systems

Johnstone [25] asked how the top eigenvalue changes in a sample covariance matrix if the population covariance matrix is not the identity, but has one (or a few) unusually large eigenvalues?

Similarly, what happens to the Tracy-Widom distribution when the mean of the entries of the GOE matrix changes? These questions have been extensively studied. In short, perturbations below a critical window do not make a difference, and above create a single unusually large eigenvalue.

For the $\beta = 2$ case, [2] derived formulas for the deformed Tracy-Widom distributions using Harish-Chandra integrals. The quest to understand the critical case for $\beta = 1$ lead to a simple derivation of the Painlevé equations for $\beta = 2, 4$ in [5].

Note that changing the mean of the GOE is just adding a rank-1 matrix. The GOE is rotationally invariant, so for eigenvalue distributions we may as well add a rank-1 perturbation of the form $e^t e$, with the first coordinate vector e. Such a perturbation commutes with tridiagonalization. At criticality, it becomes a left boundary condition for the stochastic Airy operator. The relevant theorem form Bloemendal and V. [5] is

Theorem 7.1. *Let $\mu_n \in \mathbb{R}$. Let $G = G_n$ be a $(\mu_n/\sqrt{n})$-shifted mean $n \times n$ GOE matrix. Suppose that*

$$n^{1/3}(1 - \mu_n) \to w \in (-\infty, \infty] \qquad \text{as } n \to \infty. \tag{7.1}$$

Let $\lambda_1 > \cdots > \lambda_n$ be the eigenvalues of G. Then, jointly for $k = 0, 1, \ldots$ in the sense of finite-dimensional distributions, we have

$$n^{1/6}\left(\lambda_k - 2\sqrt{n}\right) \Rightarrow -\Lambda_{k-1} \qquad \text{as } n \to \infty$$

where $\Lambda_0 < \Lambda_1 < \cdots$ are the eigenvalues of $\mathrm{SAO}_{\beta,w}$.

Here $\mathrm{SAO}_{\beta,w}$ is the Stochastic Airy operator (5.1) with left boundary condition $f'(0)/f(0) = w$. Similar theorems hold for the other β-Hermite ensembles perturbed at e.

This theorem is useful in two ways. First, it gives a characterization of the perturbed TW laws in terms of a PDE. Conversely, it gives an interpretation of the solutions of a PDE in terms of the perturbed TW laws, giving a fast way to Painlevé expressions.

Painlevé formulas. Let $u(t)$ be the Hastings-McLeod solution of the homogeneous Painlevé II equation, i.e.

$$u'' = 2u^3 + tu, \tag{7.2}$$

characterized by

$$u(t) \sim \mathrm{Ai}(t) \quad \text{as } t \to +\infty \tag{7.3}$$

where $\mathrm{Ai}(t)$ is the Airy function (6.1). Let

$$v(t) = \int_t^\infty u^2, \qquad E(t) = \exp\Big(-\int_t^\infty u\Big), \qquad F(t) = \exp\Big(-\int_t^\infty v\Big). \tag{7.4}$$

Next define two functions $f(t, w)$, $g(t, w)$ on $\mathbb{R}^2$, analytic in w for each fixed t, by the first order linear ODEs

$$\frac{\partial}{\partial w}\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} u^2 & -wu - u' \\ -wu + u' & w^2 - t - u^2 \end{pmatrix}\begin{pmatrix} f \\ g \end{pmatrix} \tag{7.5}$$

and the initial conditions

$$f(t,0) \;=\; E(t) \;=\; g(t,0). \tag{7.6}$$

Equation (7.5) is one member of the Lax pair for the Painlevé II equation. The other pair gives an ODE in the variable t. This is now sufficient information to check that $F(t,w) = f(t,w)F(t)$ satisfies the PDE (6.4), giving a proof for the Painlevé formula $P(\mathrm{TW}_2 < t) = F(t)$. However, in order to be able to check, we needed to understand where to start looking, and rank-1 perturbation theory helped!

Similar formulas hold for $\beta = 4$. For $\beta = 1$, Mo [34] has developed formulas but we do not know how to check that they satisfy the PDE.

Problem 3 (Mo's formulas). *Find a way to check that Mo's formulas satisfy* (6.4).

In [42] Rumanov finds a new (!) Painlevé representation for the hard edge using the corresponding stochastic operator. But we don't know the bulk analogue, see Question 9.

8. Beta edge universality

The transformation $(\lambda, q) \mapsto (a, b)$ in (2.4) turns complicated dependence into independence in the β-Hermite case. For more general potentials V, the first factor in (2.3) is not a product of factors depending on single variables any more, and so the variables are not independent. Still, for quartic V it can be written as a product, where each factor is a function of only two consecutive pairs (a_i, b_i).

This implies that the process $i \mapsto (a_i, b_i)$ is a Markov chain. Moreover, for general (even) polynomial V it is a η- Markov with $\eta = \deg V/2 - 1$, which means that given η consecutive pairs (a_i, b_i) the variables before and after are conditionally independent.

This observation leads naturally to a proof of universality [31]. There, it is shown that for V with $V'' > c > 0$ we have

Theorem 8.1. *There exists a coupling of the random matrices $T = T_n$ on the same probability space and constants $\gamma, \vartheta, \mathcal{E}$ depending on V only so that a.s. we have*

$$\gamma n^{2/3}(\mathcal{E}I - T_n) \to \mathrm{SAO}_\beta$$

in the norm-resolvent sense. Here $\mathcal{E}I - T_n$ acts on $\mathbb{R}^n \subset L^2(\mathbb{R}_+)$ with coordinate vectors $e_j = (\vartheta n)^{1/6}\mathbf{1}_{[j-1,j](\vartheta n)^{-1/3}}$.

Proof outline. In [38], sufficient conditions were given for the convergence of discrete operators to continuum ones, in particular to SAO_β. This was done through a more general version of the proof of Theorem 5.1.

The most important condition is that if $\mathcal{E}$ is the top edge of the equilibrium measure associated with the potential V, then the discrete version of the integrated potential converges to the continuum one, locally uniformly:

$$n^{1/3} \sum_{k=1}^{\lfloor tn^{1/3} \rfloor} (a_k + 2b_k - \mathcal{E}) \to \frac{1}{2}t^2 + \tfrac{2}{\sqrt{\beta}}b_t$$

This amounts to having to show a central limit theorem for the η-Markov chain (a_i, b_i) (we will drop the prefix η).

- The Markov chain is time-inhomogeneous because of the coefficients of the b-terms. However, these change on the scale of order n, while
- the Markov chain mixes exponentially fast, so in logarithmic number of steps it gets to its (local) stationary measure, which can be approximated using a homogeneous version of the problem.
- the local equilibrium measure is extremely close to Gaussian. Indeed, the joint distribution of stretches of length $n^{1/2-\varepsilon}$ are close in total variation to their Gaussian approximation! So the CLT is true in a very strong sense, and is proved by comparing joint densities.
- The Markov chain is *not* started from its local stationary distribution at $i = 1$. In fact, the first coordinates of the matrix T encode the local equilibrium measure for V just as they do in the β-Hermite case. Indeed, the limit of the right end of T is the Jacobi operator for the equilibrium measure associated to the potential V! See Section 3.
- Thus the CLT as required by the [38] criteria does not hold verbatim. It does hold for T truncated after the first $c \log n$ coordinates, and it can be shown that the truncation does not make a significant difference. □

By now, universality of the β-ensemble edge eigenvalues has other proofs, some more general, see [3, 7]. For the Jacobi ensembles, see [22].

Question 4 (Formulas). *There exists asymptotic formulas for correlations and other statistics of the edge and bulk processes, see for example [15]. Can these be connected to the limiting operators directly?*

Exotic edge operators. We saw in Section 3, that empirical distribution of eigenvalues of T_n, without scaling, converge to the classical equilibrium measure form potential theory corresponding to V.

The convexity and analyticity of V forces this measure to have a density which is decays like $x^{1/2}$ at the edges. As one might guess, this $x^{1/2}$ is crucial for the SAO_β limit.

When V is analytic, the possible decay rates are $x^{2k+1/2}$ for some integer k. The more detailed analysis of universality in [31] lead us to the following conjecture. See [38] for a more precise version, and a detailed explanation from where the conjectured limit comes from.

Conjecture 5. *After scaling, T_n converges to the random operator*

$$\mathcal{S}_{\beta,k} = -\partial_t^2 + t^{\frac{1}{2k+1}} + \tfrac{2}{\sqrt{\beta}}\, t^{-\frac{k}{2k+1}}\, b_t'.$$

For $\beta = 2$ the eigenvalue limits have been studied in [9] via the Riemann-Hilbert approach.

9. Bulk limits – the Brownian carousel

The goal of this section is to describe the limit of the β-Hermite ensembles in the bulk.

First, for motivation, we review some history. The nonlinear transformation $(a, b) \to (\lambda, q)$ of Section 2 is fundamental in several areas, including orthogonal polynomial theory,

the Toda lattice, and more generally, integrable systems and inverse spectral theory. It goes beyond tridiagonal matrices and point measures. A beautiful generalization, is the theory of **canonical systems**, where the correspondence is between certain matrix-valued "potentials" and measures on $\mathbb{R}$. Canonical systems are a one-parameter families of differential equations of the form

$$\lambda R_t f = K f', \qquad \text{on } [0,\eta), \qquad K = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

where R is a nonnegative definite 2×2 matrix-valued function from $[0,\eta)$, and f takes values in $\mathbb{R}^2$ on the same interval. When R is invertible everywhere, then the canonical system corresponds to the eigenvalue problem of the Dirac operator

$$R^{-1} K \partial_t \tag{9.1}$$

which is symmetric with respect to the inner product

$$\langle f, g \rangle = \int_0^\eta f_t^\dagger R_t g_t \, dt.$$

A theory canonical systems was developed by de Branges [10] in conjunction with generalizing the concept of Fourier transform.

The Hilbert-Pólya conjecture seeks to prove the Riemann hypothesis by finding a self-adjoint operator whose eigenvalues are the zeros Z of $\zeta(1/2 + iz)$ for the Riemann zeta function ζ. A famous attempt at proving the Riemann hypothesis was made by de Branges, using Dirac operators corresponding to canonical systems.

On the other hand, the Montgomery conjecture [35] claims that as $t \to \infty$, the random set $(Z - Ut)\log t$, where U is a uniform random variable on $[0,1]$, converges to the Sine_2 process, defined as the limit of eigenvalue process of the GUE in the bulk.

A natural question is whether there exists an operator (coming from canonical system) whose eigenvalues are give the Sine_2 process. The first theorem from [46] answers this in the affirmative, for all β. The operator we describe here is conjugate to a canonical Dirac operator via a Cayley transform, see [46], but the present form is more convenient for analysis.

Consider the hyperbolic Brownian motion in the Poincaré disk satisfying the SDE

$$d\mathcal{B} = \frac{1}{\sqrt{\beta(1-t)}}(1 - |B|^2) dZ \tag{9.2}$$

where Z is a complex Brownian motion with independent standard real and imaginary parts, and the time scaling corresponds to logarithmic time. Let

$$X_t = \frac{1}{\sqrt{1 - |\mathcal{B}(t)|^2}} \begin{pmatrix} 1 & \mathcal{B}(t) \\ \overline{\mathcal{B}}(t) & 1 \end{pmatrix}, \qquad J = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}. \tag{9.3}$$

Define the Brownian carousel operator as

$$\mathcal{C}_\beta = J\, X_t^2\, \partial_t \qquad \text{on } [0,1). \tag{9.4}$$

with boundary conditions $f(0)$ parallel to $(1,1)^\dagger$ and $f(1)$ parallel to $(B(1),1)^\dagger$ (since $\mathcal{B}$ converges to a point on the unit circle). We will see that $2\mathcal{C}_\beta$ has a discrete set of eigenvalues with a translation-invariant distribution. It is called the Sine_β process.

Then we have

Theorem 9.1 ([46]). *Fix $\nu \in (-2, 2)$. There exists unitary matrices so that for the β-Hermite tridiagonal matrices T_n*

$$\sqrt{1-\nu^2}\, O_n(T_n - \nu I)O_n^{-1} \to \mathcal{C}_\beta$$

where T_n acts on the $\mathbb{C}^n$ as a subspace of complex 2-vector-valued functions on $[0, 1)$. The convergence is in the norm-resolvent sense; in particular eigenvalues converge and eigenvectors converge in norm.

A version of this theorem, for unitary matrices (and for the associated phase function instead of the operator) was given Killip and Stoicu [28]. In [44] a phase function version is proved. The full operator convergence is shown in [46].

The Brownian carousel as a geometric evolution. Writing the eigenvalue equation for $\mathcal{C}_\beta$ as

$$\partial_t g = -\lambda J^{X_t^{-1}} g, \qquad g(0) = (1, 1)^\dagger.$$

Shows that $\mathcal{P}g_t = e^{i\gamma_t}$, a point on the unit circle, is rotated at speed λ about the moving center $\mathcal{PB}(t)$. In particular, γ satisfies

$$\partial_t \gamma = \lambda \frac{|e^{i\gamma} - \mathcal{B}|^2}{1 - |\mathcal{B}|^2}, \qquad \gamma(0) = 0. \tag{9.5}$$

Oscillation theory tells us that the number of eigenvalues in the interval $[0, \lambda]$ equals the number of times $e^{i\gamma}$ visits the point $\mathcal{B}(1)$. This process is called the Brownian carousel, introduced in [44] before the discovery of the operator $\mathcal{C}_\beta$.

We will not describe the proof of Theorem 9.1 here. Instead, we will explain how this operator arises as a limit of lifts of (random) unitary matrices. Then we present some applications to approximating eigenvalue statistics. Finally, we will discuss a related model, 1-dimensional critical random Schrödinger operators.

10. An operator and a path associated with unitary matrices

The goal of this section is to parameterize the spectrum of a unitary matrix in a way that it will be apparent already for finite n what the limiting operator will be. In fact, we construct a Dirac operator whose spectrum is the lifting of that of U. Moreover, the operator depends on a piecewise constant path in the hyperbolic plane. If this path has a limit as $n \to \infty$ (and some tightness conditions are satisfied) then so will the associated Dirac operator.

As it turns out, in the circular beta case the parameter path is just a random walk in the hyperbolic plane! Hence the limit will be the operator parameterized by hyperbolic Brownian motion.

The construction is based on the Szegő recursion, which we will briefly review here.

Let U be a unitary matrix of dimension n, and assume that for some unit vector e, the vectors $e, Ue, \dots U^{n-1}e$ form a basis. There is a unique way to apply Gram-Schmidt to orthonormalize this basis so that we get

$$\Phi_0(U)e, \dots, \Phi_{n-1}(U)e$$

where Φ_k is a monic degree k polynomial. Define Φ_n to be monic of degree n so that $\Phi_n(U)e = 0$; this implies that $\Phi_n(z) = \det(z - U)$ the characteristic polynomial of U. Writing

$$\Phi_k(z) = z^k + a_{k-1}z^{k-1} + \ldots + a_0$$

we define

$$\Phi_k^*(z) = \bar{a}_0 z^k + \bar{a}_1 z^{k-1} + \ldots + \bar{a}_k = z^k \overline{\Phi_k(1/\bar{z})}.$$

Now note that

$$\langle \Phi_k^*(U)e, U^j e\rangle = \sum_{i=0}^{k} \bar{a}_i \langle U^{k-i}e, U^j e\rangle = \sum_{i=0}^{k} \bar{a}_i \langle U^{k-j}e, U^i e\rangle = \overline{\langle \Phi_k(U)e, U^{k-j}e\rangle}.$$

By construction, $u = \Phi_k(U)e$ is perpendicular to $e, \ldots, U^{k-1}e$, it follows that $\Phi_k^*(U)e$ is perpendicular to $Ue, \ldots, U^k e$. However, so is $v = \Phi_{k+1}(U)e - U\Phi_k(U)e$ (as each term is, by construction). Now u, v are in the span of $e, \ldots, U^k e$, so they must be collinear. Following tradition we choose α_k, the so-called Verblunski coefficients, so that

$$\Phi_{k+1} - z\Phi_k = -\bar{\alpha}_k \Phi_k^*, \tag{10.1}$$

namely

$$-\bar{\alpha}_k = \frac{\langle u, v\rangle}{\langle u, u\rangle} = \frac{\langle \Phi_k^*(U)e, -U\Phi_k(U)e\rangle}{\|\Phi_k^*(U)e\|^2}.$$

Since $\Phi_k^*(U)e$ and $U\Phi_k(U)e$ have the same length, we see that $|\alpha_k| \leq 1$. We then get the celebrated Szegő recursion

$$\begin{pmatrix} \Phi_{k+1}(z) \\ \Phi_{k+1}^*(z) \end{pmatrix} = A_k Z \begin{pmatrix} \Phi_k(z) \\ \Phi_k^*(z) \end{pmatrix}, \qquad \Phi_0^*(z) = \Phi_0(z) = 1,$$

with the matrices

$$A_k = \begin{pmatrix} 1 & -\bar{\alpha}_k \\ -\alpha_k & 1 \end{pmatrix}, \qquad Z = \begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that z is an eigenvalue if and only if $\Phi_n(z) = 0$, equivalently by (10.1) we have

$$Z \begin{pmatrix} \Phi_{n-1}(z) \\ \Phi_{n-1}^*(z) \end{pmatrix} = ZA_{n-2}Z \cdots ZA_0 Z \begin{pmatrix} 1 \\ 1 \end{pmatrix} \parallel \begin{pmatrix} \bar{\alpha}_{n-1} \\ 1 \end{pmatrix}. \tag{10.2}$$

Using the Verblunski coefficients, we can define a new set of parameters

$$b_k = \mathcal{P} A_0^{-1} \ldots A_{k-1}^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad 0 \leq k < n-1 \tag{10.3}$$

where $\mathcal{P}\binom{x}{y} = x/y$, and

$$b_* = \mathcal{P} A_0^{-1} \ldots A_{n-2}^{-1} \begin{pmatrix} \bar{\alpha}_{n-1} \\ 1 \end{pmatrix}.$$

Then $b_0 = 0$ and the parameters $(b_1, \ldots, b_{n-1}, b_*)$ encode the same information as the α_i. This is exactly the information contained in the spectral measure $\sum_{j=1}^{n} w_j \delta_{e^{i\lambda_j}}$.

Theorem 10.1. *Consider the measure $\sum_{j=1}^n w_j \delta_{e^{i\lambda_j/n}}$ supported on n points on the unit circle, and consider the b-coordinates (10.3). For $t \in [0,1]$ let $b(t) = b_{\lfloor tn \rfloor}$, and let*

$$X_t = \frac{1}{\sqrt{1-|b(t)|^2}} \begin{pmatrix} 1 & b(t) \\ \bar{b}(t) & 1 \end{pmatrix}, \qquad J = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}.$$

Then the operator

$$JX_t^2 \partial_t \tag{10.4}$$

acting on functions $f : [0,1] \to \mathbb{C}^2$ with the boundary conditions $f_1(0) = f_2(0)$ and $f_1(1) = f_2(1)b_*$ has discrete spectrum and the eigenvalues are $\lambda_i/2 + \pi n\mathbb{Z}$.

Proof. We skip the standard proof of self-adjointness, see [46]. Instead of the Szegő recursion, we can follow the evolution of

$$\Gamma_k = Z^{A_{k-2}\dots A_0} \dots Z^{A_0} Z \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

so that

$$\Gamma_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \Gamma_1 = Z \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \Gamma_2 = Z^{A_0} Z \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \dots$$

which, geometrically is a repeated rotation of the vector around a moving center given by b_k, and

$$\Gamma_{k+1} = Z^{A_{k-1}\dots A_0} \Gamma_k = Z^{X_{k/n}^{-1}} \Gamma_k$$

Since J is an infinitesimal rotation element around 0, with $z = e^{i\lambda/n}$ the solution $\Gamma(t)$ of the ODE

$$\partial_t \Gamma(t) = -\frac{\lambda}{2} J^{X_t^{-1}} \Gamma(t), \qquad \Gamma(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

satisfies $\Gamma(k/n) = e^{-ik/2n}\Gamma_k$ for $k = 0, \dots, n$. But since $X_t J X_t^* = J$, $X_t = X_t^*$ and $J^2 = -I$, this ODE is just the eigenvalue equation at $\lambda/2$ of $JX_t^2\partial t$. Note also that $\Gamma(1)$ is parallel to the middle term of (10.2), so the boundary condition is also correct. □

11. The path parameter for circular β

We now look at the circular β ensembles. Their joint eigenvalue density is proportional to Vandermonde to the power β. What we need is that for this eigenvalue distribution we can take the α_k to be rotationally symmetric, independent with

$$|\alpha_k^2| \sim \text{Beta}\left[1, (n-k-1)\beta/2\right]$$

with α_{n-1} uniform on the circle, as shown by Killip and Nenciu [27]. The evolution of b_k is

$$b_{k+1} = A_k^{A_{k-1}\cdots A_0}.b_k$$

where the A_k are now to be understood as linear fractional transformations, or, equivalently, hyperbolic automorphisms in the Poincaré model.

Note that A_k moves the origin to a rotationally invariant random location, and so $A_k^{A_{k-1}\cdots A_0}$ moves b_k to a rotationally invariant random location around b_k. In particular,

b_k is just a random walk in the hyperbolic plane that can be described alternatively as follows. Let $b_0 = 0$. Given b_k, pick a point uniformly on the hyperbolic circle around b_k whose radius equals the hyperbolic distance d_k of 0 and a random variable with the same distribution as $|\alpha_k|$.

Given a hyperbolic Brownian motion path B, this method suggest an efficient coupling. First pick $d_1, \ldots, d_{n-1}$, let $b_0 = 0$, $t_0 = 0$, and given b_k, t_k let t_{k+1} be the first time that $\mathrm{dist}(B_t, b_k) = d_{k+1}$. Let $b_{k+1} = B(t_{k+1})$.

Given this coupling, it is now straightforward to show that the path $b_n(t) \to \mathcal{B}(t)$ a.s. uniformly on compacts, for $\mathcal{B}$ defined in (9.2). With an additional tightness argument, we get

Theorem 11.1 ([46]). *The operators $\mathcal{C}_{\beta,n}$ defined by* (10.4) *with paths b_n coupled as above, converge in the norm-resolvent sense to the limit $\mathcal{C}_\beta$ of* (9.4)*. In particular, the circular β eigenvalue process converges to the eigenvalues of C_β.*

For bulk results in the Laguerre case, see [24].

12. The Brownian carousel

The Brownian carousel description gives a simple way to analyze the limiting point process. The hyperbolic angle of the rotating boundary point as measured from $b(t)$ follows the **Brownian carousel SDE**. Indeed, define $\alpha_\lambda(t)$ to be the continuous function with $\alpha_\lambda(0) = 0$ so that with X as in (9.3) (recall $\mathcal{P}(x,y)^\dagger = x/y$)

$$e^{i\alpha_\lambda(t)} = \mathcal{P}X^{-1}g_\lambda(t)$$

for the solution g_λ of the ODE $2C_\beta g_\lambda = \lambda g_\lambda$ started at $(1,1)^\dagger$. (A factor 2 here for backward compatibility). While $\mathcal{P}g$ evolves monotonously on the circle, the evolution of α satisfies a coupled one-parameter family of stochastic differential equations. We apply a logarithmic time change for simplicity to get, with $f(t) = \frac{\beta}{4}\exp(-\beta t/4)$ the SDE

$$d\alpha_\lambda = \lambda f\, dt + \Re((e^{-i\alpha_\lambda} - 1)dZ), \qquad \alpha_\lambda(0) = 0, \tag{12.1}$$

driven by a two-dimensional standard Brownian motion. For a single λ, this reduces to the one-dimensional stochastic differential equation

$$d\alpha_\lambda = \lambda f\, dt + 2\sin(\alpha_\lambda/2)dW, \qquad \alpha_\lambda(0) = 0, \tag{12.2}$$

which converges as $t \to \infty$ to an integer multiple $\alpha_\lambda(\infty)$ of 2π. A direct consequence of oscillation theory for $\mathcal{C}_\beta$ is the following.

Proposition 12.1. *The number of points $N(\lambda)$ of the point process Sine_β in $[0,\lambda]$ has the same distribution as $\alpha_\lambda(\infty)/(2\pi)$.*

13. Gap probabilities

In the 1950s Wigner examined the asymptotic probability of having no eigenvalue in a fixed interval of size λ for $n \to \infty$ while the spectrum is rescaled to have an average eigenvalue

spacing 2π. Wigner's prediction for this probability was

$$p_\lambda = \exp\left(-(c+o(1))\lambda^2\right).$$

where this is a $\lambda \to \infty$ behavior. This rate of decay is in sharp contrast with the exponential tail for gaps between Poisson points; it is one manifestation of the more organized nature of the random eigenvalues. Wigner's estimate of the constant c, $1/(16\pi)$, later turned out to be inaccurate. [18] improved this estimate to

$$p_\lambda = (\kappa_\beta + o(1))\lambda^{\gamma_\beta} \exp\left(-\frac{\beta}{64}\lambda^2 + \left(\frac{\beta}{8} - \frac{1}{4}\right)\lambda\right) \tag{13.1}$$

which applies to the Sine_β process.

Dyson's computation of the exponent γ_β, namely $\frac{1}{4}(\frac{\beta}{2} + \frac{2}{\beta} + 6)$, was shown to be slightly incorrect. Indeed, [14] gave more substantiated predictions that γ_β is equal to $-1/8, -1/4$ and $-1/8$ for values $\beta = 1, 2$ and 4, respectively. Mathematically precise proofs for the $\beta = 1, 2$ and 4 cases were later given by several authors: [47], [13]. Moreover, the value of κ_β and higher order asymptotics were also established for these specific cases by [30], [20], [11].

In [45] we give a mathematically rigorous version of Dyson's prediction for general β with a corrected exponent γ_β using the Brownian carousel SDE.

Theorem 13.1. *The formula* (13.1) *holds with a positive* κ_β *and*

$$\gamma_\beta = \frac{1}{4}\left(\frac{\beta}{2} + \frac{2}{\beta} - 3\right).$$

We include a proof of a theorem from [44] that works for more general driving functions f (the equation (12.1)) but gives a weaker result in this case, namely the main order term in the upper bound.

Theorem 13.2. *Let* $f : \mathbb{R}^+ \to \mathbb{R}^+$ *satisfy* $f(t) \le c/(1+t^2)$ *for all* t *and* $\int_0^\infty |df| < \infty$. *Let* $k \ge 0$. *As* $\lambda \to \infty$, *for the point process given by the Brownian carousel with parameter* f *we have*

$$P(\#\text{ of points in } [0,\lambda] \le k) = \exp\big(-\lambda^2(\|f\|_2^2/8 + o(1))\big). \tag{13.2}$$

Lemma 13.3. *Let* Y *be an adapted stochastic process with* $|Y_t| < m$, *and let* X *satisfy the SDE* $dX = Y\,dB$ *where* B_t *is a Brownian motion. Then for each* $a, t > 0$ *we have*

$$P(X(t) - X(0) \ge a) \le \exp\left(-a^2/(2tm^2)\right).$$

Proof. We may assume $X(0) = 0$. Then $X_t = B_\tau$ where τ is the random time change $\tau = \int_0^t Y^2(s)ds$. Since $\tau < m^2 t$ the inequality now follows from

$$P(B_r > a) \le \exp\left(-a^2/(2r)\right). \qquad \square$$

Proof of Theorem 13.2. The event in (13.2) is given in terms of the Brownian carousel SDE as $\lim_{t\to\infty} \alpha_\lambda(t) \le 2k\pi$.

Since $\alpha(t)$ never returns below a multiple of 2π that it has passed, it is enough to give an upper bound on the probability that α stays less than $x = 2(k+1)\pi$. For $0 < s < t$ we have

$$P(\alpha(t) < x \,|\, \mathcal{F}_s) \quad = \quad P\left(-\int_s^t 2\sin(\alpha/2) dB \,>\, \lambda \int_s^t f dt - x + \alpha(s) \,\Big|\, \mathcal{F}_s \right).$$

We may drop the $\alpha(s)$ from the right hand side and use Lemma 13.3 with $Y = -2\sin(\alpha/2)$, $m = 2$, $a = \lambda(\int_s^t f dt - x/\lambda)$ to get the upper bound

$$P(\alpha(t) < x \,|\, \mathcal{F}_s) \le \exp(-\lambda^2 r(s,t)), \qquad r(s,t) = \frac{(\int_s^t f dt - x/\lambda)^2}{8(t-s)}.$$

Then, by just requiring $\alpha(t) < x$ for times $\varepsilon, 2\varepsilon, \ldots \in [0, K]$ the probability that α stays less than $x = 2(k+1)\pi$ is bounded above by

$$E \prod_{k=0}^{K/\varepsilon} P(\alpha((k+1)\varepsilon) < x \,\big|\, \mathcal{F}_{k\varepsilon}) \le \exp\Big\{ -\lambda^2 \sum_{k=0}^{K/\varepsilon} r(\varepsilon k, \varepsilon k + \varepsilon) \Big\}.$$

A choice of ε so that $x/\lambda = o(\varepsilon)$ as $\lambda \to \infty$ yields the asymptotic Riemann sum

$$\sum_{k=0}^{K/\varepsilon} r(\varepsilon k, \varepsilon k + \varepsilon) = \frac{1}{8} \int_0^K f^2(t) dt + o(1).$$

Letting $K \to \infty$ provides the desired upper bound. □

Next, we show a central limit theorem for the number of eigenvalues of $\mathcal{C}_\beta$ from [32].

Theorem 13.4 (CLT for Sine_β). *As $\lambda \to \infty$ we have*

$$\frac{1}{\sqrt{\log \lambda}} \left(\mathrm{Sine}_\beta[0, \lambda] - \frac{\lambda}{2\pi} \right) \Rightarrow \mathcal{N}(0, \frac{2}{\beta \pi^2}).$$

An $n \to \infty$ version of this theorem for finite matrices from circular and Jacobi β ensembles was shown by Killip [26].

Proof. We will consider the Brownian carousel SDE

$$d\alpha^\lambda = \lambda \frac{\beta}{4} e^{-\frac{\beta}{4}t} dt + 2\sin(\alpha^\lambda/2) dB, \quad \alpha^\lambda(0) = 0 \quad t \in [0, \infty). \tag{13.3}$$

First note that $\tilde{\alpha}(t) = \alpha^\lambda(T + t)$ with $T = \frac{4}{\beta} \log(\beta\lambda/4)$ satisfies the same SDE with $\lambda = 1$. Therefore

$$\frac{\alpha^\lambda(\infty) - \alpha^\lambda(T)}{\sqrt{\log(\lambda)}} \to 0$$

in probability. So it suffices to find the the weak limit of

$$\frac{\alpha^\lambda(T) - \lambda}{2\pi\sqrt{\log \lambda}}.$$

We have

$$\alpha(T) - \lambda = -\frac{4}{\beta} + \int_0^T 2\sin(\alpha^\lambda/2)dB$$

which means

$$\alpha(T) - \lambda + \frac{4}{\beta} \stackrel{d}{=} \hat{B}\left(\int_0^T 4\sin(\alpha^\lambda/2)^2 dt\right)$$

for a certain standard Brownian motion $\hat{B}$. In order to prove the required limit in distribution we only need to show that $\frac{4}{\log\lambda}\int_0^T \sin(\alpha^\lambda/2)^2 dt \to \frac{8}{\beta}$ in probability. We have

$$\frac{4}{\log\lambda}\int_0^T \sin(\alpha^\lambda/2)^2 dt = \frac{8\log[\beta\lambda/4]}{\beta\log\lambda} + \frac{2}{\beta\log\lambda}\int_0^T \cos(\alpha^\lambda)dt.$$

The first term converges to $8/\beta$. To bound the second term we compute

$$\begin{aligned}\frac{4}{i\beta\lambda\log\lambda} d\left(e^{i\alpha^\lambda+\beta t/4}\right) = {} & \frac{e^{i\alpha^\lambda}}{\log\lambda}dt + \frac{8}{\beta\lambda\log\lambda}e^{i\alpha^\lambda+\beta t/4}\sin(\alpha^\lambda/2)dB \\ & + \frac{8i}{\beta\lambda\log\lambda}e^{i\alpha^\lambda+\beta t/4}\sin(\alpha^\lambda/2)^2 dt \\ & + \frac{1}{i\lambda\log\lambda}e^{i\alpha^\lambda+\beta t/4}dt.\end{aligned}$$

The integral of the left hand side is $\frac{4}{i\beta\lambda\log\lambda}\left[4e^{i\alpha^\lambda(T)}\lambda/\beta - 1\right] = O((\log\lambda)^{-1})$. The integrals of the last two terms in the right hand side are of the order of $(\lambda\log\lambda)^{-1}\int_0^T e^{\beta t/4}dt = O((\log\lambda)^{-1})$. Finally, the integral of the second term on the right has an L^2 norm which is bounded by $C(\log\lambda)^{-1}$. This means the integral of the first term on the right,

$$(\log\lambda)^{-1}\int_0^T e^{i\alpha^\lambda}dt$$

converges to 0 in probability from which the statement of the theorem follows. □

14. Random Schrödinger limits

The methods developed for tridiagonal matrices also work for critical 1-dimensional random Schrödinger operators. It is interesting to compare the behavior of level statistics.

Consider the matrix

$$H_n = \begin{pmatrix} v_1 & 1 & & & & \\ 1 & v_2 & 1 & & & \\ & 1 & \ddots & \ddots & & \\ & & \ddots & \ddots & 1 & \\ & & & 1 & v_{n-1} & 1 \\ & & & & 1 & v_n \end{pmatrix} \tag{14.1}$$

where $v_k = \sigma\omega_k/\sqrt{n}$, and ω_k are independent random variables with mean 0, variance 1 and bounded third absolute moment.

To cut a long story short, one can take a limit of this operator around the global position E just as the β-Hermite models in Theorem 9.1. The resulting operator $\mathcal{S}_\tau$ is an analogue of $\mathcal{C}_\beta$, except it is driven by time-homogeneous hyperbolic Brownian motion on an interval of length $\tau = \sigma^2/(1 - E^2/4)$, which is the only parameter left in the process. In [32] we show that the large gap probabilities have a similar behaviour (exponentially decaying in the square of the gap) to the Sine_β process (see also [23] for more detailed large deviation results).

The CLT and the level repulsion are different, indicating much higher ordering. We include the geometric proof of the repulsion here, using the Brownian carousel description of Section 9. Let $\mathsf{Sch}_\tau[I]$ denote the number of eigenvalues of the operator $\tau\mathcal{S}_\tau$ in the interval I.

Theorem 14.1 (Eigenvalue repulsion, [32]). *For $\varepsilon > 0$ we have*

$$P\left\{\mathsf{Sch}_\tau[0,\varepsilon] \geq 2\right\} \leq 4\exp\left(-\frac{(\log(2\pi/\varepsilon) - \tau - 1)^2}{\tau}\right). \tag{14.2}$$

whenever the squared expression is nonnegative.

Proof. If there are at least two points in $[0,\varepsilon]$ then the Brownian carousel had to take at least one full turn. Thus

$$P\left\{\mathsf{Sch}_\tau[0,\varepsilon] \geq 2\right\} \leq P\left\{\gamma^{\varepsilon/\tau}(\tau) \geq 2\pi\right\}.$$

where γ is the solution of (9.5). From (9.5) we get

$$\gamma^{\varepsilon/\tau}(\tau) \leq \varepsilon \max_{0\leq t\leq\tau}(1 - |B_t|^2)^{-1} = \varepsilon(1 - \max_{0\leq t\leq\tau}|B_t|^2)^{-1}$$

which means that

$$\gamma^{\varepsilon/\tau}(\tau) \geq 2\pi \quad \Rightarrow \quad 1 - \frac{\varepsilon}{2\pi} \leq \max_{0\leq t\leq\tau}|B_t|^2. \tag{14.3}$$

In the Poincaré disk model the hyperbolic distance between the origin and a point z in the unit disk is given by $q(z) = \log\left(\frac{1+|z|}{1-|z|}\right)$. Thus (14.3) implies

$$\max_{0\leq t\leq\tau} q(B_t) \geq \log(2\pi/\varepsilon).$$

The probability that the hyperbolic Brownian motion leaves a ball with a large radius r in a fixed time is comparable to the probability that a one-dimensional Brownian motion leaves $[-r, r]$ in the same time. This follows by noting that Itô's formula with (9.2) gives

$$dq = \frac{dB}{\sqrt{2}} + \frac{\coth(q)}{4}dt$$

for the evolution of $q(B)$ with a standard Brownian motion B. By increasing the drift from $\coth(q)/4$ to $\infty\mathbf{1}_{q\in[0,1]} + \coth(1)/4$ we see that q is stochastically dominated by $1 + t\coth(1)/4 + |B(t)|/\sqrt{2}$ where B is standard Brownian motion and $\coth(1) < 4$. Thus

$$P\left(\max_{0\leq t\leq\tau} q(B_t) \geq \log(2\pi/\varepsilon)\right) \quad \leq \quad P\left(\max_{0\leq t\leq\tau}|B(t)| \geq \log(2\pi/\varepsilon) - 1 - \tau\right)$$

$$\leq \quad 4\exp\left(-\frac{(\log(2\pi/\varepsilon)-\tau-1)^2}{\tau}\right)$$

which proves the theorem. □

We note that continuum random Schrödinger models can also have such limits, see [29] and [36].

Most of this review was about eigenvalues. To conclude, we include a remarkable fact about the shape of localized eigenvectors of 1-dimensional random Schrödinger operators, [41].

Theorem 14.2. *Pick λ uniformly from the eigenvalues of H_n and let ψ^λ be the corresponding normalized eigenvector. Let B be a two sided Brownian motion started from 0, and let*

$$M(t) = \exp(B(t) - |t/2|).$$

Then, letting $\tau_E = \sigma^2/(1-E^2/4)/4$, as $n \to \infty$ we have the convergence in joint distribution

$$\left(\ \lambda,\ \psi^\lambda\lfloor t/n\rfloor^2 dt^*\right) \implies \left(\ E,\ M(\tau_E(t-U))dt^*\ \right)$$

where E has arcsin distribution on $[-2,2]$, U is uniform on $[0,1]$, and E, U, M are independent. Here dt^* signifies that the measures are both normalized to have total mass 1.

15. Further open problems

These are in addition to the problems and questions presented in the body of the article.

Question 6 (Decimation). *In Forrester [21] it was shown that deleting all but every kth eigenvalue of many finite $\beta = 2/k$ ensembles gives the corresponding $\beta = 2k$ ensemble. Can the limiting operators (bulk or edge) be coupled explicitly in this way?*

Question 7 (Random and deterministic orthogonal polynomials). *Is there a relation between the $\beta = 2$ random orthogonal polynomials (see section 4) and the deterministic ones? How about the limiting operators?*

Question 8 (Dynamics). *Are there operator limits of matrix-valued (say Hermitian) Brownian motion?*

Question 9 (Painlevé in the bulk). *Can one deduce the gap Painlevé equation from the PDE's corresponding to the generator of the Brownian carousel SDE?*

Question 10 (Loop equations). *Can one derive analogues of the so-called loop equations directly from limiting operators?*

Acknowledgemeents. The author is grateful to the Rényi Institute, Budapest for its hospitality during the writing of this paper.

References

[1] G. Anderson, A. Guionnet, and O. Zeitouni, *Introduction to random matrices*, Cambridge University Press, 2009.

[2] Jinho Baik, Gérard Ben Arous, and Sandrine Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Ann. Probab. **33** (2005), 1643–1697.

[3] Florent Bekerman, Alessio Figalli, and Alice Guionnet, *Transport maps for beta-matrix models and universality*, arXiv:1311.2315, 2013.

[4] Alex Bloemendal, *Finite rank perturbations of random matrices and their continuum limits*, PhD thesis, University of Toronto, 2011.

[5] Alex Bloemendal and Bálint Virág, *Limits of spiked random matrices I*, Probability Theory and Related Fields **156**(3-4) (2013), 795–825.

[6] Gaëtan Borot and Céline Nadal, *Right tail asymptotic expansion of Tracy-Widom beta laws*, Random Matrices: Theory and Applications **1**(03) (2012).

[7] Paul Bourgade, Laszlo Erdos, and Horng-Tzer Yau, *Edge universality of beta ensembles*, arXiv:1306.5728, 2013.

[8] Jonathan Breuer, Peter J Forrester, and Uzy Smilansky, *Random discrete Schrödinger operators from random matrix theory*, Journal of Physics A: Mathematical and Theoretical **40**(5) (2007), F161.

[9] Tom Claeys, Igor Krasovsky, and Alexander Its, *Higher-order analogues of the Tracy-Widom distribution and the Painlevé II hierarchy*, Communications on pure and applied mathematics **63**(3) (2010), 362–412.

[10] Louis de Branges, *Hilbert spaces of entire functions*, volume 1. Prentice-Hall Englewood Cliffs, NJ, 1968.

[11] P. Deift, A. Its, I. Krasovsky, and X. Zhou, *The Widom-Dyson constant for the gap probability in random matrix theory*, J. Comput. Appl. Math. **202**(1), (2007), 26–47.

[12] P. A. Deift, *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*, Courant Lecture Notes in Mathematics, New York, 1999.

[13] P.A. Deift, A.R. Its, and X. Zhou, *A Riemann-Hilbert approach to asymptotic problems arising in the theory of random matrices and also in the theory of integrable statistical mechanics*, Ann. Math. **146** (1997), 149–235.

[14] J. des Cloizeaux and M. L. Mehta, *Asymptotic behavior of spacing distributions for the eigenvalues of random matrices*, J. Mathematical Phys. **14** (1973), 1648–1650.

[15] Patrick Desrosiers and Dang-Zheng Liu, *Asymptotics for products of characteristic polynomials in classical beta ensembles*, Constructive Approximation, pp. 1–50, 2013.

[16] Laure Dumaz and Bálint Virág, *The right tail exponent of the Tracy-Widom-β distribution*, arXiv:1102.4818, 2011.

[17] Ioana Dumitriu and Alan Edelman, *Matrix models for beta ensembles*, J. Math. Phys. **43**(11) (2002), 5830–5847.

[18] F.J. Dyson, *Statistical theory of energy levels of complex systems II*, J. Math. Phys. **3** (1962), 157–165.

[19] Alan Edelman and Brian D Sutton, *From random matrices to stochastic operators*, Journal of Statistical Physics **127**(6) (2007), 1121–1165.

[20] Torsten Ehrhardt, *Dyson's constant in the asymptotics of the Fredholm determinant of the sine kernel*, Comm. Math. Phys. **262**(2) (2006), 317–341.

[21] P. J. Forrester, *A random matrix decimation procedure relating* $\beta = 2/(r+1)$ *to* $\beta = 2(r+1)$, ArXiv e-prints, November 2007.

[22] Diane Holcomb and Gregorio R Moreno Flores, *Edge scaling of the* β*-Jacobi ensemble*, Journal of Statistical Physics **149**(6) (2012), 1136–1160.

[23] Diane Holcomb and Benedek Valkó, *Large deviations for the sine-beta and sch-tau processes*, arXiv:1311.2981, 2013.

[24] Stéphanie Jacquot and Benedek Valkó, *Bulk scaling limit of the Laguerre ensemble*, Electronic Journal of Probability **16** (2011), 314–346.

[25] Iain M. Johnstone, *High dimensional statistical inference and random matrices*, In International Congress of Mathematicians. Vol. I, pp. 307–333. Eur. Math. Soc., Zürich, 2007.

[26] Rowan Killip, *Gaussian fluctuations for* β *ensembles*, IMRN: International Mathematics Research Notices, 2008.

[27] Rowan Killip and Irina Nenciu, *Matrix models for circular ensembles*, International Mathematics Research Notices **2004**(50) (2004), 2665–2701.

[28] Rowan Killip and Mihai Stoiciu, *Eigenvalue statistics for CMV matrices: from Poisson to clock via random matrix ensembles*, Duke Math. J. **146**(3) (2009), 361–399.

[29] S. Kotani and F. Nakano, *Level statistics of one-dimensional Schröodinger operators with random decaying potential*, ArXiv e-prints, October 2012.

[30] I. V. Krasovsky, *Gap probability in the spectrum of random matrices and asymptotics of polynomials orthogonal on an arc of the unit circle*, Int. Math. Res. Not. (25) (2004), 1249–1272.

[31] M. Krishnapur, B. Rider, and B. Virag, *Universality of the Stochastic Airy Operator*, ArXiv e-prints, June 2013.

[32] Eugene Kritchevski, Benedek Valkó, and Bálint Virág, *The scaling limit of the critical one-dimensional random Schrödinger operator*, Comm. Math. Phys. **314**(3) (2012), 775–806.

[33] Michel Ledoux and Brian Rider, *Small deviations for beta ensembles*, Electron. J. Probab **15**(41) (2010), 1319–1343.

[34] M. Y. Mo, *Rank 1 real wishart spiked model*, Communications on Pure and Applied Mathematics **65**(11) (2012), 1528–1638.

[35] Hugh L., Montgomery, *Distribution of the zeros of the Riemann zeta function*, In Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 1, pp. 379–381. Canad. Math. Congress, Montreal, Que., 1975.

[36] Fumihiko Nakano, *Level statistics for one-dimensional Schrödinger operators and Gaussian beta ensemble*, arXiv:1312.6901, 2013.

[37] José A Ramírez and Brian Rider, *Diffusion at the random matrix hard edge*, Communications in Mathematical Physics **288**(3) (2009), 887–906.

[38] José A. Ramírez, Brian Rider, and Bálint Virág, *Beta ensembles, stochastic Airy spectrum, and a diffusion*, J. Amer. Math. Soc. **24**(4) (2011), 919–944.

[39] Jose A Ramirez, Brian Rider, and Ofer Zeitouni, *Hard edge tail asymptotics*, Electronic Communications in Probability **16** (2011), 741–752.

[40] Rémi Rhodes and Vincent Vargas, *Gaussian multiplicative chaos and applications: a review*, arXiv:1305.6221, 2013.

[41] Ben Rifkind and Bálint Virág, *The shape of eigenvectors of 1d random Schrödinger operators*, In preparation., 2014+.

[42] Igor Rumanov, *Hard edge for beta-ensembles and Painlevé III*, IMRN (2013), rnt170.

[43] Hale F. Trotter, *Eigenvalue distributions of large Hermitian matrices; Wigner's semicircle law and a theorem of Kac, Murdock, and Szegő*, Adv. in Math. **54**(1) (1984), 67–82.

[44] Benedek Valkó and Bálint Virág, *Continuum limits of random matrices and the Brownian carousel*, Inventiones Math. **177** (2009), 463–508.

[45] _____, *Large gaps between random eigenvalues*, Ann. Probab. **38**(3) (2010), 1263–1279.

[46] _____, *Bulk operators*, In preparation., 2014+.

[47] H. Widom, *The asymptotics of a continuous analogue of orthogonal polynomials*, J. Approx. Theory **77** (1996), 51–64.

Department of Mathematics, University of Toronto, Canada
E-mail: balint math.toronto.edu

Constrained forms of statistical minimax: Computation, communication, and privacy

Martin J. Wainwright

Abstract. A fundamental quantity in statistical decision theory is the notion of the minimax risk associated with an estimation problem. It is based on a saddlepoint problem, in which nature plays the role of adversary in choosing the underlying problem instance, and the statistician seeks an estimator with good properties uniformly over a class of problem instances. We argue that in many modern estimation problems arising in the mathematical sciences, the classical notion of minimax risk suffers from a significant deficiency: to wit, it allows for all possible estimators, including those with prohibitive computational costs, unmanageable storage requirements, or other undesirable properties. Accordingly, we introduce some refinements of minimax risk based on imposing additional constraints on the sets of possible estimators. We illustrate this notion of constrained statistical minimax via three vignettes, based on restrictions involving computation, communication, and privacy, respectively.

Mathematics Subject Classification (2010). Primary 62Cxx; Secondary 68W40.

Keywords. Statistical minimax; information theory, metric entropy, communication complexity, computational complexity, differential privacy.

1. Introduction

Minimax theory is a cornerstone of statistical decision theory, providing a classical approach to assessing the quality of a statistical estimator in the frequentist sense. It is based on a saddle point problem, in which the adversary chooses a worst-case set of parameters, and the statistician seeks to minimize the worst-case risk via a well-chosen estimator. There is now a rich and well-developed body of theory for bounding and/or computing the minimax risk for various statistical estimation problems (e.g., see the papers [6, 26, 43, 44] and references therein).

In full generality, a statistical estimator of a parameter $\theta \in \Theta$ is a measurable function of the data, taking values in the parameter space Θ. Herein lies a serious deficiency of the classical notion of minimax risk: apart from the measurability requirement, the infimum over estimators is unconstrained. Consequently, the classical notion allows for the use of estimators that may be practically infeasible for various reasons. For instance, it allows for estimators whose computational complexity can scale arbitrarily quickly with the problem dimension and parameters. In practice, it is typically only of interest to consider estimators with polynomial-time complexity, or perhaps even more stringently, with linear or quadratic complexity. In addition, it implicitly assumes that all the data can be aggregated at a central location. For the massive data sets that are generated in many modern scientific and

Proceedings of the International Congress of Mathematicians, Seoul, 2014

engineering applications, such centralized aggregation is often impossible, and instead, distributed methods should be used. Finally, there are many types of data—including financial records, medical tests, and genetic data— that lead naturally to privacy concerns. Given the prevalence of such data types, another important issue is the study of statistical estimators that have privacy-respecting properties.

Accordingly, with the motivation of addressing these deficiencies of the classical minimax risk, the goal of this overview is to introduce and discuss various constrained forms of minimax risk. We begin in Section 2 by providing a more precise definition of the problem of statistical estimation and the notion of minimax risk. Sections 3, 4, and 5, respectively, are devoted to constrained forms of minimax risk based on communication, privacy, and computation. The results described here are based on joint pieces of work [17, 45, 47] with John Duchi, Michael Jordan, and Yuchen Zhang.

2. Classical minimax risk

In order to set the stage, we begin by describing the problem of statistical estimation in general terms, and then introducing the classical notion of minimax risk. Consider a family of probability distributions $\mathcal{P}$ with support $\mathcal{X}$, and consider a mapping $\theta : \mathcal{P} \to \Theta$. Thus, associated with member $\mathbb{P} \in \mathcal{P}$ is the parameter $\theta(\mathbb{P})$. Given a fixed but unknown distribution $\mathbb{P} \in \mathcal{P}$, suppose that we observe a sequence $X_1^n := (X_1, \ldots, X_n)$ of random variables drawn i.i.d. according to $\mathbb{P}$. Based on observing the sequence X_1^n, our goal is to estimate the *target parameter* $\theta^* := \theta(\mathbb{P})$. More formally, an estimator of θ^* is a measurable function $\widehat{\theta} : \mathcal{X}^n \to \Theta$. In order to assess the quality of any estimator, we let $\rho : \Theta \times \Theta \to [0, \infty)$ be some non-negative measure of error on the parameter space Θ, and consider the associated *risk function*

$$R(\widehat{\theta}, \theta^*) = \mathbb{E}\big[\rho(\widehat{\theta}(X_1^n), \theta^*)\big],$$

where the expectation is taken over the samples. Typical choices of the error function ρ are various metrics, or powers of such metrics.

For any fixed estimator $\widehat{\theta}$, the function $\theta^* \mapsto R(\widehat{\theta}, \theta^*)$ characterizes its performance as the underlying truth θ^* ranges over the parameter space Θ. (Here and throughout the paper, whenever the dependence of $\widehat{\theta}$ on the samples X_1^n is clear from the context, then we simply write $\widehat{\theta}$.) There are various ways in which to "scalarize" the risk function in order to assign a single number to each estimator. In the minimax setting, for each estimator $\widehat{\theta}$, we compute the worst-case risk $\sup_{\mathbb{P} \in \mathcal{P}} R(\widehat{\theta}, \theta(\mathbb{P}))$, and rank estimators according to this ordering. The estimator that is optimal in this sense defines a quantity known as the *minimax risk*—namely,

$$\mathfrak{M}_n(\theta(\mathcal{P})) := \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} R(\widehat{\theta}, \theta(\mathbb{P})), \tag{2.1}$$

where the infimum ranges over all possible estimators.

As a prelude to later results, let us consider a few illustrative instances of these definitions.

Location families: For a fixed base density function ϕ and vector $\theta \in \mathbb{R}^d$, consider a distribution $\mathbb{P}_\theta$ specified by a density function (with respect to Lebesgue measure) of the

form $f_\theta(x) = \phi(x - \theta)$. Letting Θ be some subset of $\mathbb{R}^d$, the collection of distributions $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ is known as a *location family*, since θ plays the role of a centering quantity. Important examples of location families include the normal location family (specified by the standard normal density $\phi(x) = e^{-\frac{\|x\|_2^2}{2}}/\sqrt{2\pi}$), and for $d = 1$, the uniform location family (specified by the base density $\phi(x) = \mathbb{I}[x \in [0, 1]]$, where $\mathbb{I}$ is a zero-one valued indicator function for set membership). A typical error measure is the squared ℓ_2-norm $\rho(\widehat{\theta}, \theta^*) = \|\widehat{\theta} - \theta^*\|_2^2$. We discuss the role of communication constraints for minimax rates in location families in Section 3.

Density estimation: Parameters need not be limited to vectors, but can be more general infinite-dimensional objects. As one instance, suppose that $\mathcal{P}$ consists of a family of distributions supported on the interval $[0, 1]$, and with densities with respect to Lebesgue measure. Suppose that $f^* = \theta(\mathbb{P})$ is the density of $\mathbb{P}$. In this case, an estimator $\widehat{\theta}$ returns a density function $\widehat{f}$ with support on $[0, 1]$, and a reasonable measure of error is the usual squared $L^2([0, 1])$ norm

$$\rho(\widehat{f}, f^*) = \int_0^1 \big(\widehat{f}(t) - f^*(t)\big)^2 dt. \tag{2.2}$$

We discuss this example in Section 4.

Linear regression: An instance of linear regression is specified by a known design matrix $X \in \mathbb{R}^{n \times d}$, in which each row corresponds to a vector of d predictors, and an unknown weight vector $\theta^* \in \mathbb{R}^d$. An observed response vector $Y \in \mathbb{R}^n$ is assumed to be generated by the equation

$$Y = X\theta^* + W, \tag{2.3}$$

where $W \in \mathbb{R}^n$ is a vector of i.i.d. $N(0, \sigma^2)$ variates. Equivalently, the underlying statistical model consists of the family of distributions $\{\mathbb{P}_\theta, \mid \theta \in \Theta\}$, where each $\mathbb{P}_\theta$ is the distribution of a $N(X\theta, \sigma^2 I_{n \times n})$ random vector. (This example is slightly different from our set-up, in that the components of the observed vector Y are not identically distributed for a fixed X.) One error measure is the in-sample prediction error

$$\rho_X(\widehat{\theta}, \theta^*) := \frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2. \tag{2.4}$$

The problem of linear regression under this error measure is discussed in detail in Section 5.

3. Estimation under communication constraints

Given the modern "data deluge", it is often the case that centralized methods—in which all the data can be stored on a single computer—are no longer possible to implement. Instead, distributed methods must be used. Given a cluster of m machines, it is natural to consider splitting the full data set into m separate subsets, operating separately on each subset, and then performing some sort of communication in order to agree upon a consensus estimate. In practice, the communication budget is severely limited due to power or bandwidth

constraints, and such constraints make the problem mathematically interesting. Various researchers have studied communication-efficient algorithms for statistical estimation (e.g., see the papers [2, 15, 30, 46] and references therein). In this spirit, our first vignette is devoted the role of communication constraints in statistical estimation: we define a communication-constrained version of the minimax risk, and provide sharp bounds for a few examples. See the paper [45] for further details.

Distributed estimation protocols: Recall from Section 1 the general framework of statistical estimation, based on some family of distributions $\mathcal{P}$. Suppose that, for some fixed but unknown member $\mathbb{P}$ of $\mathcal{P}$, there are $m \geq 1$ sets of data, each stored on an individual machines. For $j \in [m] := \{1, \ldots, m\}$, the j^{th} subset $X_{1,j}^n := (X_{1j}, \ldots X_{nj})$ is an i.i.d. sample of size n from the unknown distribution $\mathbb{P}$. Consequently, the *total sample size* across all machines is $N = mn$. Given this distributed collection of local data sets, our goal is to estimate $\theta(\mathbb{P})$ based on the full collection of data $X_1^N = (X_{1,j}^n, \quad j \in [m])$, but using limited communication. Of particular interest to us is the minimal number of bits that must be exchanged in order for a distributed protocol to match the centralized minimax rate—that is, the optimal performance for an estimator given direct access to all N samples.

We now define a particular class of distributed protocols Π, which operate in a sequence of rounds. At each round $t = 1, 2, \ldots$, machine j sends to a central fusion center a message $Y_{t,j}$ that is a measurable function of the local data $X_{1,j}^n$, and potentially of past messages. We use $\overline{Y}_t = \{Y_{t,j}\}_{j \in [m]}$ denote the collection of all messages sent at round t. Given a total of T rounds, the fusion center collects the sequence $(\overline{Y}_1, \ldots, \overline{Y}_T)$, and constructs an estimator $\widehat{\theta} := \widehat{\theta}(\overline{Y}_1, \ldots, \overline{Y}_T)$.

We refer to the length $L_{t,j}$ of message $Y_{t,j}$ is the minimal number of bits required to encode it, and the total length $L = \sum_{t=1}^{T} \sum_{j=1}^{m} L_{t,j}$ of all messages sent corresponds to the *total communication cost* of the protocol. Note that the communication cost is a random variable, since the length of the messages may depend on the data, and the protocol may introduce auxiliary randomness.

The simplest type of protocol is an *independent* one: it involves only on a single round ($T = 1$) of communication, in which machine j sends message $Y_{1,j}$ to the fusion center. Since there are no past messages, the message $Y_{1,j}$ is a function only of the local data $X_{1,j}^n$. Given a class of distributions $\mathcal{P}$, the class of independent protocols with budget $B \geq 0$ is given by

$$\mathcal{A}_{\mathrm{ind}}(B, \mathcal{P}) = \left\{ \text{independent protocols } \Pi \text{ such that } \quad \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\sum_{j=1}^{m} L_j \right] \leq B \right\}. \tag{3.1}$$

In the independent case, we use Y_j to indicate the message sent from processor j, and L_j to denote its length.

In contrast to independent protocols, the class of *interactive protocols* allows for interaction at different stages of the message passing process. In particular, suppose that machine j sends message $Y_{t,j}$ to the fusion center at time t, who then relays it back to all other machines in the system. This type of global broadcast system is reasonable in settings in which the processors have limited power or upstream capacity, but the centralized fusion center can send messages without limit. In the interactive setting, the message $Y_{t,j}$ should be viewed as a measurable function of the local data $X_{1,j}^n$, and the past messages $\overline{Y}_{1:t-1}$. The family of

interactive protocols with budget $B \geq 0$ is given by

$$\mathcal{A}_{\text{inter}}(B, \mathcal{P}) = \left\{ \text{interactive protocols } \Pi \text{ such that } \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[L] \leq B \right\}. \tag{3.2}$$

Distributed minimax risks: We are now equipped to define some distributed analogues of the classical minimax risk (2.1). Given a class of distributions $\mathcal{P}$, suppose that we are interested in estimating some parameter $\theta : \mathcal{P} \to \Theta$. Given a communication budget B, we apply an independent protocol Π that generates a sequence of messages $Y_1^m = (Y_1, \ldots, Y_m)$, and we use $\widehat{\theta}(Y_1^m)$ to denote an estimator that is a measurable function of these messages. With this set-up, the *minimax risk for independent protocols* under squared ℓ_2-error is given by

$$\mathfrak{M}_{n,m}^{\text{ind}}(\theta(\mathcal{P}); B) := \inf_{\Pi \in \mathcal{A}_{\text{ind}}(B, \mathcal{P})} \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}, \Pi} \big[\|\widehat{\theta}(Y_1^m) - \theta(\mathbb{P})\|_2^2 \big]. \tag{3.3}$$

Here the double infimum is taken over all independent procotols Π that satisfy the budget constraint B, and over all estimators $\widehat{\theta}(Y_1^m)$. The *minimax risk for interactive protocols*, denoted by $\mathfrak{M}_{n,m}^{\text{inter}}$, is defined analogously, where the infimum is instead taken over the class of interactive protocols.

In either case, of primary interest is the following question: how large a budget B is required so as to ensure that the distributed minimax risk (3.3) matches the classical minimax risk (2.1) up to constant factors? In the following subsections, we answer this question precisely for two different classes of statistical estimation problems.

Bounds for uniform location family: We begin by considering a univariate example, in particular the problem of estimating the location parameter in the uniform location family $\mathcal{U} = \{P_\theta, \ \theta \in [-1, 1]\}$, where $\mathbb{P}_\theta$ denotes the uniform distribution on the interval $[\theta - 1, \ \theta + 1]$.

Proposition 3.1. *Consider the uniform location family $\mathcal{U}$ with n i.i.d. observations per machine:*

(a) *There is a universal constant c such that given a budget $B = \log(1/\delta)$ for any $\delta \geq \frac{1}{mn}$, the minimax risk is lower bounded as*

$$\mathfrak{M}_{n,m}^{\text{inter}}(\theta(\mathcal{U}); B) \geq \frac{c}{\delta^2}.$$

(b) *Conversely, given a budget of $B = \big[2 + 2 \ln m\big] \log(mn)$ bits, there is a universal constant c' such that*

$$\mathfrak{M}_{n,m}^{\text{inter}}(\theta(\mathcal{U}); B) \leq \frac{c'}{(mn)^2}.$$

If each of m machines receives n observations, we have a total sample size of mn, so the minimax rate over all centralized procedures scales as $1/(mn)^2$. Consequently, Proposition 3.1 shows that the number of bits required to achieve the centralized rate has only *logarithmic* dependence on the number m of machines and local sampl size n. Part (a) shows that if $B \ll \log(mn)$, then the distributed minimax rate is larger than the centralized

optimum, so that this logarithmic scaling is unavoidable.

The proof of Proposition 3.1 is based on a somewhat more general result, one involving the geometric structure of the parameter space Θ, as captured by its metric entropy [28]. In particular, given a subset $\Theta \subset \mathbb{R}^d$, we say $\{\theta^1, \ldots, \theta^K\}$ are δ-separated if $\|\theta^i - \theta^j\|_2 > \delta$ for $i \neq j$. The *packing entropy* of Θ with respect the Euclidean norm is given by

$$\log M_\Theta(\delta) := \log_2 \big[\max \big\{ K \in \mathbb{N} \mid \{\theta_1, \ldots, \theta^K\} \subset \Theta \text{ are } \delta\text{-separated} \big\} \big]. \tag{3.4}$$

The function $\theta \mapsto \log M_\Theta(\delta)$ is left-continuous and non-increasing in δ, so we may define the inverse function $\log M_\Theta^{-1}(B) := \sup\{\delta \mid \log M_\Theta(\delta) \geq B\}$. With this notation, we have the following general result:

Theorem 3.2. *For any family of distributions $\mathcal{P}$ and parameter set $\Theta = \theta(\mathcal{P})$, the interactive minimax risk is lower bounded as*

$$\mathfrak{M}_{n,m}^{\text{inter}}(\theta(\mathcal{P}); B) \geq \Big(\frac{1}{4} \log M_\Theta^{-1}(2B + 2) \Big)^2. \tag{3.5}$$

Of course, the same lower bound also holds for $\mathfrak{M}_{n,m}^{\text{ind}}(\theta, \mathcal{P}, B)$, since any independent protocol is a special case of an interactive protocol. Theorem 3.2 is a relatively generic statement, not exploiting any particular structure of the problem; however, there are problems for which it cannot be improved by more than constant factors [45].

Bounds for Gaussian location families: Proposition 3.1 shows that achieving the minimax risk in the uniform location family requires a budget scaling only logarithmically in the number of machines m. It is natural to wonder whether such logarithmic dependence holds more generally. Here we show that it does not: for the Gaussian location family, the dependence on m must be (nearly) linear.

Consider the *d-dimensional normal location family*

$$\mathcal{N}_d([-1, 1]^d) = \{\mathcal{N}(\theta, \sigma^2 I_{d \times d}) \mid \theta \in \Theta = [-1, 1]^d\}, \tag{3.6}$$

and suppose that our goal is to estimate the mean vector $\theta \in \mathbb{R}^d$ under the error measure $\rho(\widehat{\theta}, \theta^*) = \|\widehat{\theta} - \theta^*\|_2^2$. Given a total of $N = mn$ samples, the centralized minimax rate scales as $\frac{\sigma^2}{mn}$, achieved by the sample mean. The following result addresses the minimal budget B required for a distributed protocol to match this centralized minimax rate:

Theorem 3.3. *There exists a universal (numerical) constant c such that*

$$\mathfrak{M}_{n,m}^{\text{inter}}(\mathcal{N}_d([-1, 1]^d); B) \geq c \frac{\sigma^2 d}{mn} \min \left\{ \frac{mn}{\sigma^2}, \frac{m}{\log m}, \frac{m}{(B/d + 1) \log m} \right\}. \tag{3.7}$$

Consequently, Theorem 3.3 shows that to match the classical minimax risk up to constant factors, the number of bits communicated must scale with the product of the dimension d and number of machines m—more precisely, we must have $B \asymp dm / \log m$. Apart from the logarithmic factor, this lower bound is achievable by a simple procedure: each machine computes the sample mean of its local data and quantizes each coordinate to precision σ^2/n using $\mathcal{O}(d \log(n/\sigma^2))$ bits. These quantized sample averages are communicated to the fusion center using $B = \mathcal{O}(dm \log(n/\sigma^2))$ total bits. The fusion center averages them, obtaining an estimate with mean-squared error of the optimal order $\sigma^2 d/(mn)$.

4. Minimax theory under privacy constraints

In the modern practice of statistics, privacy concerns are becoming increasingly important. Many forms of data, including financial records, medical records, and genetic tests, have associated privacy concerns. In such settings, it is natural to individuals might request some form of privacy guarantee before allowing their data to be collected. At the same time, there is a great deal of statistical utility associated with the collection of such data, including more efficient allocation of medical resources, and biomedical research into the genetic underpinnings of disease.

There is a very large body of classical research on privacy and statistical inference (e.g., [18, 19, 23, 41]. A major focus has been on the problem of reducing disclosure risk: the probability that a member of a dataset can be identified given released statistics of the dataset. In a more recent line of work, a formal definition of disclosure risk, known as differential privacy [3, 20, 21], has emerged from the theoretical computer science community, and has been the focus of considerable attention (e.g., see the papers [11, 12, 14, 22, 25, 27, 35, 42] and references therein). Here we describe how to use the notion of local differential privacy in order to define a constrained version of the minimax risk; see the paper [17] for further details.

Differential privacy: Let us begin by defining the notion of (local) differential privacy. Suppose that X_1^n represents the original data, where each X_i takes values in the space $\mathcal{X}$. As a means of preserving privacy, we release only a "privatized" sequence Z_1^n, where each Z_i takes values in the space $\mathcal{Z}$. In the case of a non-interactive mechanism, the two sequences are related via a conditional distribution $\mathbb{Q}$ that takes the product form

$$\mathbb{Q}_n(Z_1, \ldots, Z_n \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{Q}(Z_i \mid X_i). \tag{4.1}$$

We refer to $\mathbb{Q}$ as the *channel distribution*, since it acts as a conduit between the private data X and observed data Z. There are also more complicated, interactive forms of privacy mechanisms, in which the product condition (4.1) is relaxed, but we restrict attention here to this simpler case.

We now give a precise definition of local differential privacy. Let $\sigma(\mathcal{Z})$ be the σ-field over which the channel distribution $\mathbb{Q}$ is defined. Given a privacy parameter $\alpha \geq 0$, the distribution $\mathbb{Q}$ is said to satisfy α-*local-differential privacy* if

$$\sup_{S \in \sigma(\mathcal{Z})} \sup_{x, x' \in \mathcal{X}} \frac{\mathbb{Q}(Z \in S \mid X = x)}{\mathbb{Q}(Z \in S \mid X = x')} \leq \exp(\alpha), \tag{4.2}$$

This formulation of local privacy was first proposed by Evfimievski et al. [22]. Since we limit our discussion to local privacy throughout this overview, we typically omit the adjective "local" from here onwards.

The definition (4.2) has a very natural consequence in terms of disclosure risk: when the privacy parameter α is relatively close to zero, then in a uniform sense over events S, it is impossible to distinguish between two different realizations of the private variable X. Indeed, a simple argument shows that the definition (4.2) provides a lower bound on the error in a binary hypothesis test between $X = x$ and $X = x'$; see Wasserman and Zhou [42] for more details.

The *Laplace mechanism* is a simple way in which to enforce α-privacy. Given a datum X, suppose that we release the private variable $Z = X + W$, where W follows a Laplace distribution with parameter α—that is, it has density $\phi(w) = \frac{\alpha}{2} \exp\big(-\alpha|w|\big)$. In this case, for any pair $x, x' \in [0,1]$, we have

$$\frac{\mathbb{Q}(Z = z \mid X = x)}{\mathbb{Q}(Z = z \mid X = x')} = \frac{\frac{\alpha}{2}\exp\big(-\alpha|z-x|\big)}{\frac{\alpha}{2}\exp\big(-\alpha|z-x'|\big)} \leq \exp(\alpha|x-x'|) \leq \exp(\alpha), \tag{4.3}$$

showing that the Laplace mechanism provides differential privacy on the interval $[0,1]$. Part of the goal of studying the α-private minimax risk is to determine under what conditions, if any, a specific method for producing α-private variables, such as the Laplace mechanism, is optimal.

The α-private minimax risk: We now turn to a definition of the notion of an α-private minimax risk. As usual, let $\mathcal{P}$ denote a family of probability distributions on the space $\mathcal{X}$, and suppose that our goal is to estimate some parameter $\theta(\mathbb{P})$. Let $\mathcal{Q}_\alpha$ denote the class of all channel distributions $\mathbb{Q}$ satisfying α-local differential privacy (4.2). In an operational sense, any distribution $\mathbb{Q} \in \mathcal{Q}_\alpha$ can be thought of as a privacy mechanism—namely, one means of generating a privatized data set Z_1^n from the raw data X_1^n. Rather than allowing estimators to depend on the raw data, we consider only estimators $\widetilde{\theta} = \widetilde{\theta}(Z_1^n)$ that are measurable functions of the privatized data Z_1^n. For a fixed channel distribution $\mathbb{Q}$ (and hence fixed distribution over the variables Z_1^n) and a fixed estimator $\widetilde{\theta}$, the usual worst-case risk

$$\sup_{\mathbb{P}\in\mathcal{P}} R(\widetilde{\theta}(Z_1^n, \theta(\mathbb{P})) = \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P},\mathbb{Q}}\big[\rho(\widetilde{\theta}(Z_1^n, \theta(\mathbb{P})\big] \tag{4.4}$$

is a measure of the quality of $\widetilde{\theta}$. In addition to finding the optimal estimator $\widetilde{\theta}$, we also seek an *optimal privacy mechanism*—namely, a member of $\mathcal{Q}_\alpha$ for which the minimax risk is minimized. More formally, we define the *α-private minimax risk* as

$$\mathfrak{M}_n(\theta(\mathcal{P}); \alpha) := \inf_{\mathbb{Q}\in\mathcal{Q}(\alpha)} \inf_{\widetilde{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P},\mathbb{Q}}\big[\rho(\widetilde{\theta}(Z_1^n), \theta(\mathbb{P}))\big]. \tag{4.5}$$

When $\alpha = \infty$, it reduces to the usual notion of minimax risk, but of primary interest are value of α relatively close to zero. The private minimax risk (4.5) allows us to study the tradeoff between the privacy, as measured by the differential privacy parameter α, and the statistical utility, as measured by the minimax risk of all estimators that make use only of the privatized data Z_1^n.

Density estimation under α-local-privacy: We now turn to an example that demonstrates some striking differences between the ordinary and α-private minimax risks. Recall the problem of density estimation introduced in Section 1. Given n i.i.d. samples X_1^n drawn from an univariate distribution with density f^* supported on $[0,1]$, and the goal is to return an estimate $\widehat{f}$ of the unknown density function, and we evaluate its quality using the squared $L^2([0,1])$ error previously defined in equation (2.2). In this section, we state a result that demonstrates how the minimax rate for estimating density functions with Sobolev classes changes with the addition of a privacy constraint.

We begin by defining Sobolev classes in terms of elliptical subsets of the sequence space $\ell^2(\mathbb{N})$. Consider a sequence of functions $\{\phi_j\}_{j=1}^\infty$ that form an orthonormal basis for

$L^2([0,1])$, so that any function $f \in L^2([0,1])$ can be expanded as a sum $\sum_{j=1}^{\infty} \theta_j \phi_j$ in terms of the basis coefficients $\theta_j := \int f(x)\phi_j(x)dx$, and we are guaranteed that $\{\theta_j\}_{j=1}^{\infty} \in \ell^2(\mathbb{N})$. Sobolev classes are obtained by enforcing a particular decay rate on the coefficients θ. In particular, given a parameter $s \geq 1$, the generalized Sobolev class $\mathcal{F}_s([0,1])$ is given by

$$\mathcal{F}_s([0,1]) := \big\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \in L^2([0,1]) \quad \text{for a sequence } \{\theta_j\}_{j=1}^{\infty} \text{ s.t. } \sum_{j=1}^{\infty} j^{2s}\theta_j^2 \leq 1 \big\}. \tag{4.6}$$

If we choose the trigonometric basis as our orthonormal basis, then membership in the classical Sobolev class $\mathcal{F}_s([0,1])$ corresponds to certain smoothness constraints on the derivatives of f (e.g., see the book [38] for details).

In the classical (non-private) setting, the density estimate $\widehat{f}$ is constructed based on direct observation of the original samples X_1^n, where each $X_i \sim \mathbb{P}$. In this setting, it is known [36, 38] that the minimax risk for non-private estimation of densities in the class $\mathcal{F}_s([0,1])$ scales as

$$\mathfrak{M}_n\big(\mathcal{F}_s([0,1])\big) \asymp \big(\frac{1}{n}\big)^{\frac{2s}{2s+1}}. \tag{4.7}$$

For instance, when $s = 1$, corresponding to Lipschitz functions when using the trigonometric basis, then the minimax rate scales as $n^{-\frac{2}{3}}$. Naturally, the minimax rate increases towards the parametric rate n^{-1} as the smoothness parameter s tends to infinity. The minimax rate (4.7) can be achieved by various methods, with one of the simplest being the orthogonal series estimator. Given the samples X_1^n, this method is based on computing the empirical basis coefficients $\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^n \phi_j(X_i)$, and then setting

$$\widehat{f} = \sum_{j=1}^{T} \widehat{\theta}_j \phi_j, \qquad \text{where } T = n^{\frac{1}{2s+1}}. \tag{4.8}$$

The specified choice of truncation level T provides the optimal trade-off between the bias and variance of the estimator, and some calculations show that it achieves the minimax rate (4.7), assuming that the smoothness level s is known to the method.

Now consider the case of α-private density estimation, in which we only observe a privatized version Z_1^n of the raw data X_1^n. The following theorem [17] characterizes the minimax rate when the α-private channel is chosen in an optimal way:

Theorem 4.1. *Consider the Sobolev class $\mathcal{F}_s([0,1])$ of densities for some $s \geq 1$. Then there are universal constants $0 < c_\ell \leq c_u < \infty$ such that for all $\alpha \in (0, 1/4]$, the (non-interactive) α-private minimax rate* (4.5) *is sandwiched as*

$$c_\ell \big(\frac{1}{n\alpha^2}\big)^{-\frac{2s}{2s+2}} \;\leq\; \mathfrak{M}_n\left(\mathcal{F}_s([0,1]); \alpha\right) \;\leq\; c_u \big(\frac{1}{n\alpha^2}\big)^{-\frac{2s}{2s+2}}. \tag{4.9}$$

The private minimax rate (4.9) differs from the classical rate (4.7) in two key ways. The effective sample size is reduced from n to $\alpha^2 n$, and more importantly, the exponent is reduced from $\frac{2s}{2s+1}$ to $\frac{2s}{2s+2}$. Thus, in the case of Lipschitz densities ($s = 1$), the minimax rate changes from $n^{-\frac{2}{3}}$ to $n^{-\frac{1}{2}}$. Consequently, Theorem 4.1 reveals a fundamental tradeoff

between privacy and statistical utility for density estimation.

How is the α-private minimax rate (4.9) achieved? In order to answer this question, two choices must be made: a choice of the α-private channel that generates the privatized samples Z_1^n, and an estimator that takes the private data as input. It is natural to wonder whether the α-private Laplace mechanism (4.3)—namely, forming the samples $Z_i = X_i + W_i$ where W_i is α-Laplace noise—combined with the orthogonal series estimate might achieve the optimal private rate. Interestingly, this approach turns out to be highly sub-optimal, as can be established by recourse to known results on nonparametric deconvolution. Given the observation $Z = X + W$, the density of Z is a convolution of the densities of X and W. In their study of the deconvolution problem, Carroll and Hall [10] show that if the additive noise has a characteristic function ϕ_W with tails behaving as $|\phi_W(t)| = \mathcal{O}(|t|^{-a})$ for some $a > 0$, then no method can estimate the s-smooth density of X to accuracy greater than $n^{-2s/(2s+2a+1)}$. Note that the Laplace distribution has a characteristic function with tails decaying as t^{-2}; consequently, as a special case of this result, no estimator based on applying the Laplace mechanism directly to the samples can attain a rate of convergence better than $n^{-2s/(2s+5)}$.

An optimal mechanism for α-private density estimation: This cautionary calculation motivates consideration of privacy mechanisms that are not simply based on direct perturbation of the samples X_1^n, and here we describe one such mechanism that achieves the α-private minimax rate (4.9). Recall the truncation level $T = n^{\frac{1}{2s+1}}$ from our earlier discussion of the orthogonal series estimator (4.8). Now consider the T-dimensional vectors

$$\phi(X_i) = \begin{bmatrix} \phi_1(X_i) & \phi_2(X_i) & \cdots & \phi_T(X_i) \end{bmatrix}, \tag{4.10}$$

defined for each $i = 1, \ldots, n$. These vectors are sufficient statistics for computing the orthogonal series estimator. Accordingly, our goal is to construct a channel $\mathbb{Q}$ with output space $\mathcal{Z} = \mathbb{R}^T$ such that

$$\mathbb{E}[Z_i \mid X_i] = \phi(X_i) \quad \text{for each } i = 1, \ldots, n. \tag{4.11}$$

Our construction assumes that the orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ is b_0-uniformly bounded, meaning that $\sup_x |\phi_j(x)| \leq b_0 < \infty$ for all $j = 1, 2, \ldots$. Note that many standard bases, among them the trigonometric basis and the Walsh basis, satisfy this type of boundedness condition. For some fixed $b > b_0$ to be specified, the following privacy mechanism takes as input any T-dimensional vector of the form $\tau = \phi(X_i)$ for $i = 1, \ldots, n$, as previously defined in equation (4.10). It consists of three steps, and returns a vector $Z_i \in \mathbb{R}^T$ that is α-private, and such that the unbiasedness condition (4.11) holds.

- Given a vector τ in the box $[-b_0, b_0]^T$, form a random vector $\widetilde{\tau} \in \{-b_0, b_0\}^T$ with independently sampled coordinates

$$\widetilde{\tau}_j = \begin{cases} +b_0 & \text{with probability } \frac{1}{2} + \frac{\tau_j}{2b_0}. \\ -b_0 & \text{otherwise.} \end{cases}$$

- Draw a Bernoulli random variable V equal to 1 with probability $e^{\alpha}/(e^{\alpha}+1)$, and then

draw $Z_i \in \{-b, b\}^T$ according to

$$Z_i \sim \begin{cases} \text{Uniform on } \left\{z \in \{-b,b\}^T \mid \langle z, \widetilde{\tau}\rangle > 0\right\} & \text{if } V = 1 \\ \text{Uniform on } \left\{z \in \{-b,b\}^T \mid \langle z, \widetilde{\tau}\rangle \leq 0\right\} & \text{if } V = 0. \end{cases} \tag{4.12}$$

It can be shown that the random vector Z_i is α-differentially private for any initial vector in the box $[-b_0, b_0]^T$, and moreover, the samples (4.12) are efficiently computable, say by rejection sampling. Iteration of expectation yields

$$\mathbb{E}[Z_i \mid X = x] = c_T \frac{b}{b_0\sqrt{T}} \left(\frac{e^\alpha}{e^\alpha + 1} - \frac{1}{e^\alpha + 1}\right) \phi(x) = c_T \frac{b}{b_0\sqrt{T}} \frac{e^\alpha - 1}{e^\alpha + 1} \phi(x), \tag{4.13}$$

for a constant c_T bounded away from zero. Consequently, setting $b = \frac{b_0\sqrt{T}}{c_T}\frac{e^\alpha+1}{e^\alpha-1}$ ensures that the unbiasedness condition (4.11) is satisfied.

Based on this α-private mechanism, we can compute the T-dimensional random vector $\widetilde{\theta} := \frac{1}{n}\sum_{i=1}^n Z_i$, which is guaranteed to be an unbiased estimate of the vector $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n \phi(X_i)$ of empirical basis coefficients. Using this unbiased estimate, we can then form the density estimate $\widetilde{f} = \sum_{j=1}^T \widetilde{\theta}_j \phi_j$. As shown in the paper [17], this estimator achieves the α-private minimax rate (4.9).

5. Computationally-constrained minimax rates

In the classical definition of minimax risk (2.1), the infimum is allowed to range over all measurable functions $\widehat{\theta} : \mathcal{X}^n \to \mathbb{R}^n$. In practice, however, one is limited to estimators with computational complexity that scales polynomially in the problem parameters. For this reason, it is natural to consider more refined notions of minimax rate, in which constraints are imposed on the computational complexity of the underlying estimators. For many problems, at least up to constant factors, the classical minimax risk can be achieved by computationally efficient estimators. In these cases, the computationally constrained minimax risk is no different than the classical minimax risk. Thus, such refinements of minimax rates are interesting only when it is possible to establish a gap between the performance of optimal procedures, and that of computationally constrained methods. A recent line of work [4, 29] has established such gaps for testing problems involving sparse and low-rank matrices, working under a conjecture in average-case complexity theory. Here we describe how, under a standard assumption in worst-case complexity theory, such a gap exists for the problem of high-dimensional sparse regression [47].

High-dimensional sparse regression: We begin by describing the problem of sparse regression and discussing some possible estimators, both computationally efficient and inefficient ones. As previously discussed, linear regression is a canonical problem in statistics, in which a response vector $Y \in \mathbb{R}^n$ is related to matrix $X \in \mathbb{R}^{n \times d}$ of covariates via the observation model (2.3). Given the goal of estimating θ^*, the quality of an estimate $\widehat{\theta}$ can be assessed in various ways. In this discussion, we model the matrix X as a fixed quantity, known as the case of deterministic design, and consider the (in-sample) *prediction error*, as previously defined in equation (2.4).

Recent years have witnessed intense study of the sparse form of linear regression, in which the unknown regression vector $\theta^* \in \mathbb{R}^d$ is assumed to have at most $k \ll d$ non-zero entries (e.g., see the papers [5, 9, 16, 24, 31, 34, 39, 40] and references therein). The most direct approach to solving a k-sparse instance of the linear regression model (2.3) is to seek a k-sparse minimizer to the least-squares cost $\|Y - X\theta\|_2^2$. Doing so leads to the ℓ_0-based estimator

$$\widehat{\theta}_{\ell_0} := \arg\min_{\theta \in \mathbb{B}_0(k)} \|Y - X\theta\|_2^2. \tag{5.1}$$

Note that this estimator returns an estimate that belongs to the ℓ_0-"ball"

$$\mathbb{B}_0(k) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} \mathbb{I}[\theta_j \neq 0] \leq k\big\} \tag{5.2}$$

of k-sparse vectors. More generally, given an estimator $\widetilde{\theta}$, we say that it belongs to class $\mathcal{A}(k)$ if its output always belongs to $\mathbb{B}_0(k)$.

The following result [8, 34] provides an upper bound on the prediction error performance of the ℓ_0-based estimator:

Proposition 5.1 (Prediction error for $\widehat{\theta}_{\ell_0}$). *There is a universal constant c_0 such that for any design matrix X, the ℓ_0-based estimator $\widehat{\theta}_{\ell_0}$ satisfies*

$$\max_{\theta^* \in \mathbb{B}_0(k)} \mathbb{E}\Big[\frac{1}{n}\|X(\widehat{\theta}_{\ell_0} - \theta^*)\|_2^2\Big] \leq c_0 \frac{\sigma^2 k \log d}{n}. \tag{5.3}$$

Moreover, Raskutti et al. [34] establish a lower bound that is matching up to constant factors, showing that this bound is unimprovable when $k \ll d$. A notable aspect of the upper bound (5.3) is that it holds for any design matrix $X \in \mathbb{R}^{n \times d}$.

Thus, in terms of the classical minimax risk (2.1), the ℓ_0-based estimator is an optimal method. However, it is unattractive from a computational point of view. A brute force approach requires iterating over all $\binom{d}{k}$ subsets of size k, and Natarajan [32] shows that computing a sparse solution to a set of linear equations is an NP-hard problem. Given this intractability, it is natural to consider the performance of computationally efficient methods.

Prediction guarantees for ℓ_1-based methods: Convex relaxation is a standard method for replacing a combinatorial constraint—in this case, the requirement that θ have at most k non-zero entries—with a looser but convex constraint. A familiar relaxation of the ℓ_0-constraint is to replace it with an ℓ_1-norm. For concreteness, we consider a Lagrangian form of the ℓ_1-relaxation, which leads to the *Lasso estimator* [13, 37]

$$\widehat{\theta}_{\ell_1} = \arg\min_{\theta \in \mathbb{R}^d} \big\{\frac{1}{2n}\|Y - X\theta\|_2^2 + \lambda_n \|\theta\|_1\big\}. \tag{5.4}$$

In contrast to the ℓ_0-based estimator (5.1), it is easy to compute the Lasso estimate. Indeed, a quadratic program of the form (5.4) can be solved to δ-accuracy in time polynomial in the problem parameters, and $\log(1/\delta)$, by various standard optimization methods (e.g., see the books [7, 33]).

Is the Lasso estimator (5.4) is an optimal method? For some error metrics, including the ℓ_2-norm error $\|\widehat{\theta} - \theta^*\|_2$, it can be shown that the Lasso is essentially an optimal method, in

terms of matching the classical minimax risk [34]. However, for the prediction error (2.4), the best known results on the Lasso fail to match the ℓ_0-guarantee. In particular, in contrast to the ℓ_0-estimate (5.1), the best known results on either the Lasso [5], or the closely related Dantzig selector [9], all involve constraints known as (sparse) restricted eigenvalue (RE) conditions, which we define next.

Restricted eigenvalues are defined in terms of subsets S of the index set $\{1, 2, \ldots, d\}$, and a cone associated with any such subset. In particular, letting S^c denote the complement of S, we define the cone $\mathbb{C}(S) := \{\theta \in \mathbb{R}^d \mid \|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1\}$. Here $\|\theta_{S^c}\|_1 := \sum_{j \in S^c} |\theta_j|$ corresponds to the ℓ_1-norm of the coefficients indexed by S^c, with $\|\theta_S\|_1$ defined similarly.

Definition 5.2 (Sparse restricted eigenvalue). The matrix $X \in \mathbb{R}^{n \times d}$ is said to satisfy a *uniform γ-RE* condition if

$$\frac{1}{n}\|X\theta\|_2^2 \geq \gamma\|\theta\|_2^2 \qquad \text{for all } \theta \in \bigcup_{|S|=k} \mathbb{C}(S). \tag{5.5}$$

The restricted eigenvalue constant of X, denoted by $\gamma(X)$, is the greatest γ such that X satisfies the condition (5.5). The RE condition (5.5) and related quantities have been studied extensively in past work on basis pursuit and the Lasso (e.g., [5, 9, 31, 34]). van de Geer and Buhlmann [39] provide an overview of the different types of RE parameters, and their relationships. The following result [5] provides an upper bound on the Lasso prediction error under a sparse RE condition:

Proposition 5.3 (Prediction error for Lasso). *There is a universal constant c_1 such that for any column-normalized design matrix X with a RE constant $\gamma(X) > 0$, the Lasso estimate $\widehat{\theta}_{\ell_1}$ satisfies*

$$\max_{\theta^* \in \mathbb{B}_0(k)} \mathbb{E}\Big[\frac{1}{n}\|X(\widehat{\theta}_{\ell_1} - \theta^*)\|_2^2\Big] \leq \frac{c_1}{\gamma^2(X)} \frac{\sigma^2 k \log d}{n}. \tag{5.6}$$

Apart from the difference in universal constants, the key difference between the ℓ_1-guarantee and the ℓ_0-guarantee is that the RE constant $\gamma^2(X)$ appears in the Lasso bound (5.6), but is absent from the ℓ_0-bound (5.3). It is natural to wonder whether it might be possible to prove a sharper bound on the Lasso, not involving the RE constant. From a fundamental point of view, given the goal of minimizing the prediction risk (2.4), the restricted eigenvalues of X should not be relevant. For instance, duplicating two rows of X would force the RE constant to zero, but would not make the underlying prediction problem any more difficult. We are thus left with two possibilities:

- either the analysis leading to the bound (5.6) is not sharp, and could be improved;
- or the appearance of the RE constant is unavoidable for an ℓ_1-based method.

Our recent work shows that in fact, the second option is correct, and even more generally, the appearance of the RE constant is intrinsic to the class of polynomial-time estimators.

Computationally-constrained minimax risk: In order to state our main result, we need to make precise a particular notion of a polynomial-efficient estimator. Since the observation (X, Y) consists of real numbers, any efficient algorithm can only take a finite-length

representation of the input. Consequently, we begin by introducing an appropriate notion of discretization. For any input value x and integer τ, the operator

$$\lfloor x \rfloor_\tau := 2^{-\tau} \lfloor 2^\tau x \rfloor \tag{5.7}$$

represents a $2^{-\tau}$-precise quantization of x. (Here $\lfloor u \rfloor$ denotes the largest integer smaller than or equal to u.) Given a real value x, an efficient estimator is allowed to take $\lfloor x \rfloor_\tau$ as its input for some finite choice τ. We denote by $\mathrm{size}(x; \tau)$ the length of binary representation of $\lfloor x \rfloor_\tau$, and denote by $\mathrm{size}(X, Y; \tau)$ the total length of the discretized matrix vector pair (X, Y).

The following definition of computational efficiency is parameterized in terms of three quantities: (i) a positive integer b, corresponding to the number of bits required to implement the estimator as a computer program; (ii) a polynomial function G of the triplet (n, d, k), corresponding to the discretization accuracy of the input, and (iii) a polynomial function H of input size, corresponding to the runtime of the program.

Definition 5.4 (Polynomial-efficient estimators). Given a pair of polynomial functions G: $(\mathbb{Z}_+)^3 \to \mathbb{R}_+$, $H : \mathbb{Z}_+ \to \mathbb{R}_+$ and a positive integer $b \in \mathbb{Z}_+$, an estimator $(Y, X) \mapsto \widehat{\theta}(Y, X)$ is said to be (b, G, H)-*efficient* if:

- It can be represented by a computer program that is encoded in b bits.
- For every problem of scale (n, d, k), it accepts inputs quantized to accuracy $\lfloor \cdot \rfloor_\tau$ where the quantization level is bounded as $\tau \leq G(n, d, k)$.
- For every input (X, Y), it is guaranteed to terminate in time $H(\mathrm{size}(X, Y; \tau))$.

In computational complexity theory, the class **POLY** corresponds to problems that are solvable in polynomial time by a Turing machine. A closely related class denoted by **PPOLY**, corresponds to all problems solvable in polynomial time by a Turing machine with a so-called advice string—meaning a side-input to the machine—that is of polynomial length. The class **PPOLY** is strictly bigger than the class **POLY** (e.g, [1]); however, it is widely believed that $\mathbf{NP} \not\subset \mathbf{PPOLY}$, and the following result is stated using this inclusion as an assumption. Moreover, we use $c_j, j = 2, 3$ to denote universal constants independent of the scaling parameters (n, d, k), polynomials (F, G, H) and constants (γ, σ, δ).

Theorem 5.5. *If* $\mathbf{NP} \not\subset \mathbf{PPOLY}$*, then for any positive integer* b*, any scalar* $\delta \in (0, 1)$*, any polynomial functions* $G : (\mathbb{Z}_+)^3 \to \mathbb{R}_+$ *and* $F, H : \mathbb{Z}_+ \to \mathbb{R}_+$*, there is a sparsity level* $k \geq 1$ *such that the following holds:*

For any dimension $d \in [4k, F(k)]$*, any sample size* n *in the interval* $[c_2 k \log d, F(k)]$*, and any scalar* $\gamma \in [2^{-G(n,d,k)}, \frac{1}{24\sqrt{2}})$*, there is a matrix* $X \in \mathbb{R}^{n \times d}$ *such that:*

(a) *It has an RE constant* $\gamma(X)$ *that is bounded as* $|\gamma(X) - \gamma| \leq 2^{-G(n,d,k)}$.

(b) *For any* (b, G, H)*-efficient estimator* $\widehat{\theta} \in \mathcal{A}(k)$*, the mean-squared prediction error is lower bounded as*

$$\max_{\theta^* \in \mathbb{B}_0(k)} \mathbb{E}\Big[\frac{\|X(\widehat{\theta} - \theta^*)\|_2^2}{n}\Big] \geq \frac{c_3}{\gamma^2} \, \frac{\sigma^2 k^{1-\delta} \log d}{n}. \tag{5.8}$$

Disregarding technical aspects regarding quantization, the essential part of the theorem is that the lower bound grows inversely with the squared RE constant γ^2. Consequently, within the class of polynomial-time methods, the Lasso is an optimal method, but faster rates can be obtained using algorithms with exponential-time complexity. We note that Theorem 5.5 is restricted to methods that return k-sparse estimates—that is, belong to the class $\mathcal{A}(k)$. It is an open question as to whether analogous lower bounds can be established without this requirement.

Acknowledgements. Research partially supported by National Science Foundation grant CIF-31712-23800, and Office of Naval Research MURI grant N00014-11-1-0688. This lecture is based on pieces of joint work with John Duchi, Michael Jordan and Yuchen Zhang.

References

[1] S. Arora and B. Barak, *Computational Complexity: A Modern Approach*, Cambridge University Press, 2009.

[2] M. F. Balcan, A. Blum, S. Fine, and Y. Mansour, *Distributed learning, communication complexity and privacy*, http://arxiv.org/abs/1204.3514, 2012.

[3] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, *Privacy, accuracy, and consistency too: A holistic solution to contingency table release*, In Proceedings of the 26th ACM Symposium on Principles of Database Systems, 2007.

[4] Q. Berthet and P. Rigollet, *Computational lower bounds for sparse PCA*, Technical report, Princeton University, April 2013. arxiv1304.0828.

[5] P. Bickel, Y. Ritov, and A. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Annals of Statistics **37**(4) (2009), 1705–1732.

[6] L. Birgé, *Approximation dans les espaces metriques et theorie de l'estimation*, Z. Wahrsch. verw. Gebiete **65** (1983), 181–327.

[7] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, UK, 2004.

[8] F. Bunea, A. Tsybakov, and M. Wegkamp, *Aggregation for Gaussian regression*, Annals of Statistics **35**(4) (2007), 1674–1697.

[9] E. J. Candés and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n*, Annals of Statistics **35**(6) (2007), 2313–2351.

[10] R. Carroll and P. Hall, *Optimal rates of convergence for deconvolving a density*, Journal of the American Statistical Association **83**(404) (1988), 1184–1186.

[11] K. Chaudhuri and D. Hsu, *Convergence rates for differentially private statistical estimation*, In Proceedings of the 29th International Conference on Machine Learning, 2012.

[12] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, *Differentially private empirical risk minimization*, Journal of Machine Learning Research **12** (2011), 1069–1109.

[13] S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Computing **20**(1) (1998), 33–61.

[14] A. De, *Lower bounds in differential privacy*, In Proceedings of the Ninth Theory of Cryptography Conference, 2012.

[15] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao, *Optimal distributed online prediction using mini-batches*, Journal of Machine Learning Research **13** (2012), 165–202.

[16] D. L. Donoho, *Compressed sensing*, IEEE Trans. Info. Theory **52**(4) (April 2006), 1289–1306.

[17] J. C. Duchi, M. J. Wainwright, and M. I. Jordan, *Local privacy and minimax bounds: Sharp rates for probability estimation*, Technical report, UC Berkeley, 2013.

[18] G. T. Duncan and D. Lambert, *Disclosure-limited data dissemination*, Journal of the American Statistical Association **81**(393) (1986), 10–18.

[19] ———, *The risk of disclosure for microdata*, Journal of Business and Economic Statistics **7**(2) (1989), 207–217.

[20] C. Dwork, *Differential privacy: a survey of results*, In Theory and Applications of Models of Computation, volume 4978 of Lecture Notes in Computer Science, pp. 1–19, Springer, 2008.

[21] C. Dwork, F. McSherry, K. Nissim, and A. Smith, *Calibrating noise to sensitivity in private data analysis*, In Proceedings of the 3rd Theory of Cryptography Conference, pp. 265–284, 2006.

[22] A. V. Evfimievski, J. Gehrke, and R. Srikant, *Limiting privacy breaches in privacy preserving data mining*, In Proceedings of the Twenty-Second Symposium on Principles of Database Systems, pp. 211–222, 2003.

[23] I. P. Fellegi, *On the question of statistical confidentiality*, Journal of the American Statistical Association **67**(337) (1972), 7–18.

[24] E. Greenshtein and Y. Ritov, *Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization*, Bernoulli **10** (2004), 971–988.

[25] M. Hardt and K. Talwar, *On the geometry of differential privacy*, In Proceedings of the Fourty-Second Annual ACM Symposium on the Theory of Computing, pp. 705–714, 2010.

[26] R. Z. Has'minskii, *A lower bound on the risks of nonparametric estimates of densities in the uniform metric*, Theory Prob. Appl. **23** (1978), 794–798.

[27] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, *What can we learn privately?*, SIAM Journal on Computing **40**(3) (2011), 793–826.

[28] A. Kolmogorov and B. Tikhomirov, *ϵ-entropy and ϵ-capacity of sets in functional spaces*, Uspekhi Mat. Nauk. **86** (1959), 3–86. Appeared in English as Amer. Math. Soc. Translations **17** (1961), 277–364.

[29] Z. Ma and Y. Wu, *Computational barriers in minimax submatrix detection*, preprint arXiv:1309.5914, 2013.

[30] R. McDonald, K. Hall, and G. Mann, *Distributed training strategies for the structured perceptron*, In North American Chapter of the Association for Computational Linguistics (NAACL), 2010.

[31] N. Meinshausen and B. Yu, *Lasso-type recovery of sparse representations for high-dimensional data*, Annals of Statistics **37**(1) (2009), 246–270.

[32] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Computing **24**(2) (1995), 227–234.

[33] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer Academic Publishers, New York, 2004.

[34] G. Raskutti, M. J. Wainwright, and B. Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, IEEE Trans. Information Theory **57**(10) (Octorber 2011), 6976–6994.

[35] A. Smith, *Privacy-preserving statistical estimation with optimal convergence rates*, In Proceedings of the Fourty-Third Annual ACM Symposium on the Theory of Computing, 2011.

[36] C. J. Stone, *Optimal global rates of convergence for non-parametric regression*, Annals of Statistics **10**(4) (1982), 1040–1053.

[37] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society, Series B **58**(1) (1996), 267–288.

[38] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer, New York, 2009.

[39] S. van de Geer and P. Buhlmann, *On the conditions used to prove oracle results for the Lasso*, Electronic Journal of Statistics **3** (2009), 1360–1392.

[40] M. J. Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)*, IEEE Trans. Information Theory **55** (May 2009), 2183–2202.

[41] S. Warner, *Randomized response: a survey technique for eliminating evasive answer bias*, Journal of the American Statistical Association, **60**(309) (1965), 63–69.

[42] L. Wasserman and S. Zhou, *A statistical framework for differential privacy*, Journal of the American Statistical Association **105**(489) (2010), 375–389.

[43] Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Annals of Statistics **27**(5) (1999), 1564–1599.

[44] B. Yu, *Assouad, Fano and Le Cam*, In Festschrift for Lucien Le Cam, pp. 423–435, Springer-Verlag, Berlin, 1997.

[45] Y. Zhang, J. C. Duchi, , M. I. Jordan, and M. J. Wainwright, *Information-theoretic lower bounds for distributed statistical estimation with communication constraints*, Technical report, UC Berkeley, 2013. Presented at the NIPS Conference 2013.

[46] Y. Zhang, J. C. Duchi, and M. J. Wainwright, *Communication-efficient algorithms for statistical optimization*, Journal of Machine Learning Research **14** (November 2013), 3321–3363.

[47] Y. Zhang, M. J. Wainwright, and M. I. Jordan, *Lower bounds on the performance of polynomial-time algorithms for sparse linear regression*, Technical report, UC Berkeley, 2014. arXiv:1402.1918.

Department of Statistics University of California, Berkeley Berkeley, California 94720 USA
E-mail: wainwrig@berkeley.edu

13. Combinatorics

Coloring graphs with forbidden induced subgraphs

Maria Chudnovsky

Abstract. Since graph-coloring is an NP-complete problem in general, it is natural to ask how the complexity changes if the input graph is known not to contain a certain induced subgraph H. Results of Kamínski and Lozin, Holyer, and Levin and Galil imply that the problem remains NP-complete, unless H is the disjoint union of paths. Recently, the question of coloring graphs that do not contain certain induced paths has received considerable attention. Only one case of that problem remains open for k-coloring when $k \geq 4$, and that is the case of 4-coloring graphs with no induced 6-vertex path. However, little is known for 3-coloring. In this paper we survey known results on the topic, and discuss recent developments.

Mathematics Subject Classification (2010). Primary 05C15; Secondary 05C85.

Keywords. Graph coloring, induced subgraphs, coloring algorithms.

1. Introduction

Let G be a graph. We denote by $V(G)$ the vertex set of G, and by $E(G)$ the edge set of G. For a positive integer k, a *k-coloring* of G is a function $c : V(G) \to \{1, \ldots, k\}$ such that $c(u) \neq c(v)$ for every adjacent pair of vertices u, v; if such a function exists, we say that G *admits a k-coloring* or *is k-colorable*. The *chromatic number* $\chi(G)$ of a graph G is the smallest number k for which G admits a k-coloring. The algorithmic problem of determining the chromatic number of a graph is notoriously difficult; in fact it was one of the initial problems Karp showed to be NP-complete [14]. The algorithmic question remains difficult if we fix the parameter $k \geq 3$ (as opposed to allowing it to be part of the input), and ask to determine whether a given graph is k-colorable. This problem (knows as *the k-coloring problem*) was shown to be NP-complete in [18]. (Determining if a graph is 2-colorable is easy). It is therefore of interest to establish classes of graphs in which the k-coloring problem can be solved efficiently (i.e. in time that is a polynomial function of the size of the input graph). In this paper we focus on classes of graphs defined by forbidding certain substructures, called induced subgraphs.

2. Making the problem easier

In this paper when we say that an algorithm runs "in polynomial time" or is a "polynomial-time algorithm" we always mean "time polynomial as a function of the number of vertices

Proceedings of the International Congress of Mathematicians, Seoul, 2014

in the input graph"; we will not concern ourselves with the exact polynomial function that provides a bound on the complexity.

A *cycle* C_t is a graph with vertices $c_1, \ldots, c_t$ such that c_i is adjacent to c_j if and only if $|i-j| =\in \{1, t-1\}$, and a path P_t is a graph with vertices $p_1, \ldots, p_t$ such that p_i is adjacent to p_j if and only if $|i-j| = 1$. Let G be a graph. An *induced subgraph* of G is a graph H such that $V(H) \subseteq V(G)$, and $uv \in E(H)$ if and only if $uv \in E(G)$ for all $u, v \in V(H)$. Given graphs G and F we say that G *contains* F if F is isomorphic to an induced subgraph of G. G is *F-free* if G does not contain F. The question then becomes: for which graphs F and integers k can the k-coloring problem be solved in polynomial time for F-free graphs?

The discussion in the remainder of this section assumes that $P \neq NP$. Unfortunately, the news is not good on this front, because of the following powerful result of Kaminski and Lozin [13]:

Theorem 2.1. *For every $k, g \geq 3$, the problem of k-coloring graphs with no cycles of length at most g is NP-complete.*

Applying 2.1 with $g = |V(H)|$, we then obtain the following:

Theorem 2.2. *Let H be a graph with a cycle. For every $k \geq 3$, the problem of k-coloring H-free graphs is NP-complete.*

In other words, if the k-coloring problem is polynomial-time solvable for the class of H-free graphs (where $k \geq 3$), then H is a forest. It turns out that further restrictions need to be placed on H before excluding H as an induced subgraph may impact the complexity of k-coloring. A *k-edge-coloring* of a graph G is a function $c : E(G) \to \{1, \ldots, k\}$ such that $c(e) \neq c(f)$ for every pair e, f of edges of G that share and end. The algorithmic problem of k-edge-coloring is NP-complete for all fixed $k \geq 3$ [12, 15].

Next consider a construction. The *line graph* of a graph G, denoted by $L(G)$, is the graph with vertex set $E(G)$, and e and f are adjacent in $L(G)$ if and only if they share an end in G. Clearly k-edge-coloring G and k-coloring $L(G)$ are equivalent problems. A *claw* is the graph with four vertices x, y, z, w, and three edges xy, xz, xw. It is an easy exercise to show that line graphs are claw-free. We therefore deduce:

Theorem 2.3. *For every $k \geq 3$, the problem of k-coloring claw-free graphs is NP-complete.*

Consequently,

Theorem 2.4. *Let H be a graph that contains a claw. For every $k \geq 3$, the problem of k-coloring H-free graphs is NP-complete.*

A *component* of a graph is its maximal connected subgraph. Together 2.2 and 2.4 imply the following:

Theorem 2.5. *Let $k \geq 3$ be an integer, and H be a graph. If the problem of determining whether an H-free graph is k-colorable can be solved in polynomial time, then every component of H is a path.*

The remainder of this paper deals with coloring H-free graphs, where H is a path.

3. Excluding induced paths

In this section we summarize what is currently known about the complexity of coloring graphs with certain induced paths excluded. It turns out that excluding short induced paths does in fact help with coloring. P_2-free graphs have no edges, P_3-free graphs are disjoint unions of complete graphs, and P_4-free graphs with at least two vertices are either not connected or not connected in the complement (by a theorem of Seinsche [17]); in all cases polynomial-time coloring algorithms can easily be constructed. The first non-trivial case is the class of P_5-free graphs:

Theorem 3.1 ([10]). *For every integer k, the k-coloring problem can be solved in polynomial time for the class of P_5-free graphs.*

On the other hand,

Theorem 3.2 ([11]).

(1) *The 5-coloring problem is NP-complete for the class of P_6-free graphs.*

(2) *The 4-coloring problem is NP-complete for the class of P_7-free graphs.*

This immediately implies that:

Theorem 3.3 ([11]).

(1) *The k-coloring problem is NP-complete for the class of P_t-free graphs for all $k \geq 5$ and $t \geq 6$.*

(2) *The 4-coloring problem is NP-complete for the class of P_t-free graphs for all $t \geq 7$.*

To deduce 3.3 from 3.2, observe that for every integer $t \geq 0$ a P_t-free graph is also P_{t+1}-free, and further note that if G is a P_t-free graph that is a difficult instance for k-coloring (where $k \geq 3$ and $t \geq 2$), then the graph obtained from G by adding a new vertex adjacent to all members of $V(G)$ is a P_t-free graph that is a difficult instance for $k+1$-coloring.

Finally, in [16] it is shown that:

Theorem 3.4 ([16]). *The 3-coloring problem can be solved in polynomial time for the class of P_6-free graphs.*

Our focus here is on a new result of [4, 5]:

Theorem 3.5 ([4, 5]). *The 3-coloring problem can be solved in polynomial time for the class of P_7-free graphs.*

Thus, at the time of the writing of this manuscript, the following cases remain open:

1. the complexity of 3-coloring P_t-free graphs where $t \geq 8$, and
2. the complexity of 4-coloring P_6-free graphs.

Recently there has been some progress on the second question:

Theorem 3.6 ([11]). *The 4-coloring problem can be solved in polynomial time for the class of graphs that are both P_6-free and C_4-free.*

and

Theorem 3.7 ([6]). *The 4-coloring problem can be solved in polynomial time for the class of graphs that are both P_6-free and C_5-free.*

There are many similarities between the proofs of 3.5 and 3.7, and we will discuss them both.

4. List coloring

Given a graph G and a function L from $V(G)$ to the set of all subsets of positive integers, a *coloring* of (G, L) is a function $c : V(G) \to \bigcup_{v \in V(G)} L(v)$ such that $c(u) \neq c(v)$ for every adjacent pair uv of vertices of G, and $c(v) \in L(v)$ for every $v \in V(G)$. In this case c is also called *list coloring*, and $L(v)$ is called *the list of* v. We say that (G, L) is *colorable* if such a coloring exists. Clearly k-coloring is an instance of list coloring, where $L(v) = \{1, \ldots, k\}$ for all $v \in V(G)$.

Since list coloring is a generalization of coloring, it is an NP-complete problem in general. However, a few special cases can be solved in polynomial time, and we make use of this fact in our work. The first such case is the following:

Theorem 4.1 ([8]). *There is a polynomial-time algorithm with the following specifications:*

Input: *A pair (G, L) such that $|L(v)| \leq 2$ for all $v \in V(G)$.*

Output: *A coloring of (G, L), or a determination that none exists.*

The proof of 4.1 consists of a reduction to a well-known polynomial-time solvable problem called *2-SAT* that we do not describe here. In fact, a slightly more general result can be obtained using similar techniques (a set of vertices S is monochromatic in a given coloring c if $c(v) = c(u)$ for all $u, v \in S$.):

Theorem 4.2 ([1]). *There is a polynomial-time algorithm with the following specifications:*
Input:

(1) *a pair (G, L) such that $|L(v)| \leq 2$ for all $v \in V(G)$, and*

(2) *a list $S_1, \ldots, S_{|V(G)|^t}$ of subsets of $V(G)$, where t is an integer.*

Output: *A coloring of (G, L), so that each of $S_1, \ldots, S_{|V(G)|^t}$ is monochromatic, or a determination that none exists.*

Here is another, much more difficult, result from [1]

Theorem 4.3 ([1]). *There is a polynomial-time algorithm with the following specifications:*

Input: *A pair (G, L) such that G is P_6-free and $L(v) \subseteq \{1, 2, 3\}$ for all $v \in V(G)$.*

Output: *A coloring of (G, L), or a determination that none exists.*

Armed with these three theorems, our strategy for both 3.5 and 3.7 is to reduce the original problem of k-coloring a graph G to a polynomial number of problems $(G_1, L_1), \ldots, (G_t, L_t)$, such that

1. G is k-colorable if and only if there exists $i \in \{1, \ldots, t\}$ such that (G_i, L_i) is colorable,
2. each of $(G_1, L_1), \ldots, (G_t, L_t)$ can be solved efficiently using 4.1, 4.2 or 4.3, and
3. a k-coloring of G can be reconstructed from a coloring of (G_i, L_i) in polynomial time.

We remark that the following natural extension of 3.5 remains open:

Question 4.4. *Given a pair (G, L) where G is a P_7-free graph, and $L(v) \subseteq \{1, 2, 3\}$ for every $v \in V(G)$, what is the complexity of determining whether (G, L) is colorable?*

The corresponding extension of 3.7 is open as well. In the next section we will explain where our methods fall short for attacking these questions.

The only positive result we can report in this direction is the following:

Theorem 4.5 ([3]). *There is a polynomial-time algorithm with the following specifications:*

Input: *A pair (G, L) such that G is a bipartite P_7-free graph and $L(v) \subseteq \{1, 2, 3\}$ for all $v \in V(G)$.*

Output: *A coloring of (G, L) or a determination that none exists.*

5. Tools for 3-coloring

A *dominating set* in a graph G is a subset $S \subseteq V(G)$ such that every vertex of $V(G) \setminus S$ has a neighbor in S. For a subset X of $V(G)$, we denote by $G|X$ the subgraph of G induced by X. For $v \in V(G)$, we denote by $N(v)$ the set of vertices of G adjacent to v (in particular, $v \notin N(v)$).

In view of 4.1, the following seems like a natural approach to constructing the algorithm of 3.5:

1. Prove that there exists an integer K such that every connected P_7-free graph has a dominating set of size at most K.
2. Find a dominating set S of size at most K in G.
3. Let c be a coloring of $G|S$. Set $L(v) = \{c(v)\}$ for every $v \in S$, and

$$L(v) = \{1, 2, 3\} \setminus \bigcup_{u \in N(v)} \{c(u)\}$$

for every $v \in V(G) \setminus S$.

Clearly this is a polynomial-time procedure that reduces the original problem to at most $(3|V(G)|)^K$ instances (G, L), where each instance can be efficiently solved using 4.1.

Unfortunately, the first statement above is not true as stated, but the following questions is of interest:

Question 5.1. *Which P_7-free graphs G have a dominating set of size at most $\log |V(G)|$?*

So we cannot follow the approach outlined above directly. Instead, we identify a number of efficiently detectable "reducible configurations" in the input graph, making the graph simpler without changing its colorability properties, until a small dominating set emerges (possibly considering a number of different candidates for being a dominating set).

Next we list some examples of reducible configurations. For a graph G, a set $X \subseteq V(G)$ and a vertex $v \in V(G) \setminus X$, we say that v is *complete* to X if v is adjacent to every vertex of X, and that v is *anticomplete* to X if v has no neighbor in X.

Dominating vertex. $u, v \in V(G)$ such that $N(u) \subseteq N(v)$. Please note that it follows that u and v are non-adjacent. Clearly in this case G is 3-colorable if and only if $G \setminus u$ is 3-colorable (by making u and v be the same color). We remark that this simple reduction is the first obstacle we encounter when trying to apply our methods to 4.4, for in the list coloring setting we would additionally need to require that $L(v) \subseteq L(u)$, which is not an inherent structural property of a graph, and thus is harder to impose.

Homogeneous pair of stable sets. A *stable set* is a set of vertices all pairwise non-adjacent. Let $A, B \subseteq V(G)$ be disjoint and non-empty. We say that (A, B) is a *homogeneous pair* in G if every vertex of $V(G) \setminus (A \cup B)$ with a neighbor in A is complete to A, and the same for B. If in addition both A and B are stable sets, then (A, B) is a *homogeneous pair of stable sets*. Let (A, B) be a homogeneous pair of stable sets such that there is at least one edge between A and B, and suppose that $|A| + |B| \geq 3$. Let $a \in A$ be adjacent to $b \in B$, and let $G' = G \setminus ((A \setminus \{a\}) \cup (B \setminus \{b\}))$. It is now easy to see that G is 3-colorable if and only if G' is.

Connected neighborhood. Let $v \in V(G)$ be such that the graph $N = G|N(v)$ is connected. If N is not 2-colorable, then clearly G is not 3-colorable. Since we can check in polynomial time if a graph is 2-colorable, we may assume that N is bipartite, and, since it is connected, it has a unique 2-coloring; let N_1, N_2 be the color classes. Now, in every 3-coloring of G, the sets N_1 and N_2 are monochromatic. Let G' be obtained from $G \setminus v$ by, for $i = 1, 2$, replacing each N_i by a new vertex n_i adjacent to $(\bigcup_{n \in N_i} N(n)) \setminus \{v\}$. Now G' is P_7-free, and G is 3-colorable if and only if G' is.

The algorithm of 3.5 consists of two main parts. The first part deals with triangle-free P_7-free graphs, and exploits the structural information that follows from these assumptions. The second part deals with $P7$-free graphs that contain a triangle. Here the structure is more complex, but, on the other hand, if such a graph does have a 3-coloring, some of it can easily be seen to be forced, which makes the analysis simpler. In the next two sections we discuss the two parts of the algorithm.

6. 3-coloring P_7-free graphs: the triangle-free case

The goal of this section is to describe the algorithm of 3.5 for the case when the input is a triangle-free graph. Let G be a graph. If P is an induced path with vertices $p_1, \ldots, p_t$ in G, where $p_i p_j \in E(G)$ if and only if $|j - i| = 1$, we write "$p_1 - \ldots - p_t$ is an induced path in G" (or "is a P_t in G"). Similarly, if C is an induced cycle with vertices $c_1, \ldots, c_t$ in G, where $c_i c_j \in E(G)$ if and only if $|j - i| \in \{1, t - 1\}$, we write "$c_1 - \ldots - c_t - c_1$ is an indcued cycle in G" (or "is a C_t in G"). The *complement* G^c of G is the graph with vertex set $V(G)$ and such that $uv \in E(G^c)$ if and only if $uv \notin E(G)$. A *clique* in G is a set of vertices all pairwise adjacent. The *clique number* of G, denoted by $\omega(G)$, is the largest size of a clique in G. G is called *perfect* if $\omega(H) = \chi(H)$ for every induced subgraph H of G. Clearly, for any fixed integer k, a perfect graph G is k-colorable if and only if G does not contain a clique of size $k + 1$, and so testing if a perfect graph is k-colorable can be done in polynomial time, simply by examining all subsets of size $k + 1$. (In fact, a much stronger statement and deeper result is true: one can find the chromatic number of a perfect graph in polynomial time [9], but we do not need this here.)

The following is a well-known structural fact about perfect graphs, the Strong Perfect Graph Theorem:

Theorem 6.1 ([7]). *A graph G is perfect if and only if no induced subgraph of G or G^c is isomorphic to C_{2n+1} with $n \geq 2$.*

By 6.1, if a triangle-free P_7-free graph is not perfect, then it contains either a C_5 or a C_7. Thus in this case it is easy to check if the input graph is perfect (as with coloring, a much harder theorem is that testing perfection can be done in polynomial time [2], but we do not need it), and, in view of the discussion in this first paragraph of the section, we may assume that it is not, for otherwise we are done.

Let us thus assume that the input graph G is connected, P_7-free, triangle-free, and contains a C_7, say $c_1 - c_2 - c_3 - c_4 - c_5 - c_6 - c_7 - c_1$ (the case of C_5 uses similar ideas); write $C = \{c_1, \ldots, c_7\}$. For $i \geq 1$, let X_i be the set of vertices at distance i from C, thus the vertices of X_1 have neighbors in C, the vertices of X_2 are anticomplete to C but have neighbors in X_1, etc. Since G is triangle-free, it is easy to check that for every $x \in X_1$ there exist $p, q, r \in C$, such that $x - p - q - r$ is an induced path in G, and therefore, $X_i = \emptyset$ for $i \geq 4$. For every $x \in X_1$, let $P(x)$ denote some such path $x - p - q - r$. We may assume that G contains no reducible configurations.

Next we show that $X_3 = \emptyset$ as well. Suppose first that there exist $x, y \in X_3$ adjacent to each other. Let $x_2 \in X_2$ be adjacent to x. Since G is triangle-free, $x_2 y \notin E(G)$. By the definition of X_2, there is $x_1 \in X_1$ adjacent to x_2. But now combining $y - x - x_2 - x_1$ with $P(x_1)$ we obtain an induced seven-vertex path in G, a contradiction. Thus X_3 is a stable set. Next, let $x_3 \in X_3$. Then $N(x_3) \subseteq X_2$ and $N(x_3)$ is a stable set, since G is triangle-free. Let $x_2 \in N(x_3)$, and let $x_1 \in X_1$ be a neighbor of x_2. If x_1 is complete to $N(x_3)$, then $N(x_3) \subseteq N(x_1)$, and G contains a reducible configuration, a contradiction. So x_1 has a non-neighbor $x_2' \in N(x_3)$. But now combining $x_1 - x_2 - x_3 - x_2'$ with $P(x_1)$ we obtain an induced seven-vertex path in G, a contradiction. This proves that $X_3 = \emptyset$.

Next let us analyze the structure of X_2. First observe that if $G|X_2$ contains an odd cycle D (which therefore has length at least 5), then for every $x_1 \in X_1$ with a neighbor in D, there exist $u, v, w \in V(D)$ such that $x_1 - u - v - w$ is an induced path in G, and combining this path with $P(x_1)$, we obtain an induced seven-vertex path in G, a contradiction. This proves that $G|X_2$ is bipartite. Let F be a connected component of $G|X_2$, and suppose that $|V(F)| > 1$. Let (A, B) be a bipartition of F. We claim that (A, B) is a homogeneous pair of stable sets in G. Suppose not; then we may assume that there exist $x_1 \in X_1$ and $a_1, a_2 \in A$ such that x_1 is adjacent to a_1 and not to a_2. Choosing a_1 and a_2 with this property and subject to that at minimum distance in F, we may assume that there is $b \in B$ adjacent to both a_1 and a_2. Since G is triangle-free, $x_1 - a_1 - b - a_2$ is an induced path in G, and combining it with $P(x_1)$ we obtain an induced seven-vertex path in G, a contradiction. This proves the claim that (A, B) is a homogeneous pair of stable sets in G, and since G contains no reducible configurations, we deduce that $|A| = |B| = 1$. To summarize, we showed that every component of $G|X_2$ consists either of a single vertex, or of two adjacent vertices. Let $d_1, \ldots, d_k$ be the vertices of the singleton component, and let $\{a_1, b_1\}, \ldots, \{a_m, b_m\}$ be the vertex sets of the components of size two.

At this point we recall the outline of the algorithm we described at the start of Section 5. In our current set-up, C is not a dominating set of G, however, the set of vertices of $V(G) \setminus C$ that are anticomplete to C (namely X_2) is well under control. We proceed by considering all possible colorings of C, and assigning lists of size one to vertices of C, lists of size at most

two to vertices of X_1, and lists $\{1,2,3\}$ to vertices of X_2. Thus we have so far replaced our original problem by at most 3^7 list-coloring instances, each of which can "almost" be handled by 4.1: only vertices of X_2 may have lists of size three, but we know a lot about the structure of X_2. The rest of the proof consists of removing the "almost" in the previous sentence. We will not be able to explain this completely here, but let us show a few more steps.

To deal with $d_1, \ldots, d_k$, we "guess" (by examining all possibilities) a few (constantly many) vertices with certain properties and their colors, thus creating a set dominating all of $d_1, \ldots, d_k$. This allows us to reduce the sizes of the lists of $d_1, \ldots, d_k$ to two. Please note that while until now we only guessed colors of certain fixed vertices, thus branching into a constant number of list-coloring problems, at this stage we also guess the vertices that we pre-color, which creates polynomially many sub-problems.

Now we deal with $\{a_1, b_1\}, \ldots, \{a_m, b_m\}$. For simplicity, let us assume that every two vertices of X_1 have a common neighbor in C. Justifying this assumption requires additional arguments, but we will skip them here, referring the reader to [4]. Let X be the set of vertices of X_1 that have neighbors in $\{a_1, b_1, \ldots, a_m, b_m\}$. Let $V = \{v_1, \ldots, v_m\}$ be a set of new vertices. We now construct a new bipartite graph H, with bipartition (X, V) in which $x \in X$ is adjacent to $v_i \in V$ if and only if x has a neighbor in $\{a_i, b_i\}$. Suppose that for some $x, y \in X$ and $i, j \in \{1, \ldots, m\}$ we have $xv_i, yv_j \in E(H)$ and $xv_j, yv_i \notin E(H)$. We may assume that, in G, x is adjacent to a_i, and y to a_j. Now choosing $c \in C$ to be a common neighbor of x and y, we obtain that $b_i - a_i - x - c - y - a_j - b_j$ is a seven-vertex path in G, a contradiction. This proves that no such $x, y \in X$ and $i, j \in \{1, \ldots, m\}$ exist, which implies that H has a very special structure. In particular, it is not difficult to see that some vertex $x_0 \in X$ is complete in H to V. In G this means (up to renaming some vertices) that x_0 is complete to $\{a_1, \ldots, a_m\}$. We can now find such a vertex x_0 in polynomial time, and examine all possibilities for the color of x_0 (by branching into sub-problems). Fixing the color of x_0, in turn, allows us to reduce the size of $L(a_i)$ to two for all $i \in \{1, \ldots, m\}$. At this point, in each of the sub-problems we are considering, only $b_1, \ldots, b_m$ have lists of size three. These lists can be dealt with similarly to those of $d_1, \ldots, d_k$ by "guessing" and pre-coloring a few more "important" vertices, thus arriving at a situation where each sub-problem can be handled by 4.1.

7. 3-coloring P_7-free graphs: using triangles

The goal of this section is to discuss the algorithm of 3.5 when the input graph contains a triangle. The main idea here is to take advantage of the simple fact that all three-colorings of a triangle are the same (up to permuting colors), and, moreover, starting with the coloring of a triangle, the colors of certain other vertices are forced.

Let G be a P_7-free graph that contains a triangle. A *tripod* in G is a triple (A_1, A_2, A_3) of pairwise disjoint subsets of $V(G)$ such that

- $A_1 \cup A_2 \cup A_3 = \{v_1, \ldots, v_t\}$, where $t \geq 3$
- $v_i \in A_i$ for $i \in \{1, 2, 3\}$
- $v_1v_2v_3$ is a triangle
- Let $i \in \{1, 2, 3\}$, $\{j, k\} = \{1, 2, 3\} \setminus \{i\}$, and $p \in \{4, \ldots, t\}$. If $v_p \in A_i$, then v_p has a neighbor in $\{v_1, \ldots, v_{p-1}\} \cap A_j$ and a neighbor in $\{v_1, \ldots, v_{p-1}\} \cap A_k$.

The first step of the algorithm is to choose a maximal tripod (A_1, A_2, A_3). It is easy to see that in every 3-coloring of G, each of the sets A_1, A_2, A_3 is monochromatic, thus if one of A_1, A_2, A_3 is not a stable set, the algorithm stops and outputs a determination that no 3-coloring exists. Next we observe that, by the maximality of the tripod, every $v \in V(G) \setminus (A_1 \cup A_2 \cup A_3)$ has neighbors in at most one of A_1, A_2, A_3. Let X_i be the set of all $v \in V(G) \setminus (A_1 \cup A_2 \cup A_3)$ with a neighbor in A_i. As in the previous section, classify the remaining vertices of G by their distance from $A_1 \cup A_2 \cup A_3$. Let Y_i be the set of vertices at distance i (so $Y_1 = X_1 \cup X_2 \cup X_3$; Y_2 is anticomplete to $A_1 \cup A_2 \cup A_3$, but every vertex of Y_2 has a neighbor in $X_1 \cup X_2 \cup X_3$; etc). Observe that for every $i \in \{1, 2, 3\}$ and $x \in X_i$, and $j \in \{1, 2, 3\} \setminus \{i\}$, there is $a \in A_i$ and $b \in A_j$ such that $x - a - b$ is an induced path in G. This implies that $Y_k = \emptyset$ for every $i \geq 5$.

Once again we may assume that G has no reducible configuration. Applying arguments similar to those in the previous section (using also the connected neighborhood reducible configuration this time), we further deduce that $Y_4 = \emptyset$.

Next we prove that there exists $S \subseteq Y_1 \cup Y_2 \cup Y_3$, such that $|S| \leq 100$ and "almost" all vertices of $Y_2 \cup Y_3$ have neighbors in S. Ignoring the "almost" qualification, we are now done: since the coloring of $A_1 \cup A_2 \cup A_3$ is unique up to permuting colors, there are only constantly many possible 3-colorings of $Z = A_1 \cup A_2 \cup A_3 \cup S$, and Z is ("almost") a dominating set in G. Thus we can analyze all possible colorings of Z, obtaining a new list-coloring problem from each of them, and each of these new problems can be solved using 4.1. As in the previous section, in order to complete the proof, we guess a few more vertices that need to be added to Z to create a dominating set in G, thus branching into polynomially many sub-problems.

8. 4-coloring

In this section we briefly discuss the ideas of the proof of 3.7, some of which may extend to the more general question of 4-coloring P_6-free graphs. Let G be a P_6-free graph; then G contains no induced cycles of length at least seven. We may assume that $\omega(G) \leq 4$, and that G does not contain C_9^c, for otherwise G is not 4-colorable. This implies that G^c contains no odd cycle of length at least nine. As discussed in Section 6, we may assume that G is not perfect, and so by 6.1, G contains either a C_5 or a C_7^c.

From now on we restrict our attention to the C_5-free case, which is the subject of 3.7. Let C be the vertex set of a C_7^c in G, let X be the set of vertices of G with a neighbor in C, and $Y = V(G) \setminus (C \cup X)$. It is now easy to check that vertices in X come in two "flavors":

- $x \in X$ such that $N(x) \cap C$ contains a triangle; we call such vertices *major*, and
- $x \in X$ for which there exist $a, b, c \in C$ such that $x - a - b - c$ is a path in G; we call such vertices *minor*.

As in the previous two sections, our strategy here is to analyze all possible colorings of C. Having fixed a coloring of C, we update the lists of the vertices of X; let L be the function describing the lists. Then $|L(x)| = 1$ for every major vertex x, and $|L(x)| \in \{2, 3\}$ for every minor vertex x. Let us from now on ignore the existence of vertices in X with lists of size three (in fact, in [6] they are treated similarly to vertices of Y, rather than of X).

A *clique cutset* of a graph G is a clique K of G such that $G \setminus K$ is not connected. Clique cutsets can be readily used in coloring algorithms [19], due to the fact the $G|K$ has a unique

coloring (up to permuting colors). Now, if for some component D of $G|Y$, all vertices of X with a neighbor in D are major, then we can treat G in a way similar to a graph with a clique cutset (because G has a cutset with a unique coloring).

Let us next discuss minor vertices. It is an easy fact that if $x \in X$ is minor, and D is a component of $G|Y$, then x is either complete or anticomplete to $V(D)$. An *anticomponent* of a graph is a maximal connected induced subgraph H of G such that H^c is connected (thus H^c is a component of G^c). A more difficult, but very useful observation is that if X' is the set of all minor vertices of X with a neighbor in Y, and A is an anticomponent of $G|X'$, then $L(a) = L(b)$ for all $a, b \in V(A)$.

Let D be a component of $G|Y$, and let $A \neq \emptyset$ be the set of minor vertices of X with a neighbor in D. Then A is complete to D. To illustrate our approach, let us assume that D contains a triangle. Then A is a stable set (since $\omega(G) \leq 4$), and in particular, $G^c|A$ is connected. We may therefore assume that $L(a) = \{1, 2\}$ for all $a \in A$. Moreover, A is monochromatic in every 4-coloring of G.

For every $d \in D$, let

$$L_1(d) = \{2, 3, 4\} \setminus \bigcup_{x \in N(d), \text{ and } x \text{ is major}} L(x),$$

and

$$L_2(d) = \{1, 3, 4\} \setminus \bigcup_{x \in N(d), \text{ and } x \text{ is major}} L(x).$$

Now each of the problems (D, L_1) and (D, L_2) can be solved efficiently by 4.3.

Next consider $G' = G \setminus D$. For $i = 1, 2$, if (D, L_i) is not colorable, let $L'(a) = L(a) \setminus \{i\}$ for all $a \in A$. For $v \in V(G') \setminus A$, let $L'(v) = L(v)$. Now (G, L) is colorable if and only if (G', L') is colorable, with the additional constraint that the set A is monochromatic. Applying similar arguments to the remaining components of Y, we reduce the original problem to a problem that can be solved efficiently using 4.2.

9. Open problems

In this section, for the reader's convenience, we repeat the open problems mentioned earlier in the paper.

Question 9.1 (first mentioned as 4.4). *Given a pair (G, L) where G is a P_7-free graph, and $L(v) \subseteq \{1, 2, 3\}$ for every $v \in V(G)$, what is the complexity of determining whether (G, L) is colorable?*

Question 9.2 (first mentioned as 5.1). *Which P_7-free graphs G have a dominating set of size at most $\log |V(G)|$? What about a constant size dominating set?*

Question 9.3 (first mentioned in Section 3). *What is the complexity of 3-coloring P_t-free graphs where $t \geq 8$?*

Question 9.4 (first mentioned in Section 3). *What is the complexity of 4-coloring P_6-free graphs?*

Acknowledgments. Partially supported by NSF grants IIS-1117631 and DMS-1265803. This paper is based on the author's joint work with Peter Maceli, Juraj Stacho and Mingxian Zhong. The author is grateful to them, and to Irena Penev, for their careful reading of the present manuscript, and for many helpful suggestions. Juraj Stacho's help in surveying previously known results was invaluable. The author also thanks Alex Scott for telling her about the beautiful questions that motivated the research described here.

References

[1] H. Broersma, F. Fomin, P. Golovach, and D. Paulusma, *Three complexity results on coloring P_k-free graphs*, European journal of combinatorics **34** (3), 609–619.

[2] M. Chudnovsky, G. Cornuéjols, X. Liu, P. Seymour, and K. Vušković, *Recognizing Berge graphs*, Combinatorica **25** (2005), 143–187.

[3] M. Chudnovsky and P. Maceli, unpublished.

[4] M. Chudnovsky, P. Maceli, and M. Zhong, *Three-coloring graphs with no induced seven-vertex path I : the triangle-free case*, manuscript.

[5] _____, *Three-coloring graphs with no induced seven-vertex path II : using a triangle*, in preparation.

[6] M. Chudnovsky, P. Maceli, J. Stacho, and M. Zhong, *Four-coloring graphs with no induced six-vertex path and no induced five-wheel*, in preparation.

[7] M. Chudnovsky, N.Robertson, P. Seymour, and R.Thomas, *The strong perfect graph theorem*, Annals of Math **164** (2006), 51–229.

[8] K. Edwards, *The complexity of colouring problems on dense graphs*, Theoret. Comput. Sci. **43** (1986), 337–343.

[9] M. Gröstchel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer Verlag, 1988.

[10] C.T. Hoàng, M. Kamiński, V.V. Lozin, J. Sawada, and X. Shu, *Deciding k-colorability of P_5-free graphs in polynomial time*, Algorithmica **57** (2010), 74–81.

[11] S. Huang, *Improved Complexity Results on k-Coloring P_t-Free Graphs*, Proc. MFCS 2013, LNCS, to appear.

[12] I. Holyer, *The NP-completeness of edge coloring*, SIAM J. Comput. **10** (1981), 718–720.

[13] M. Kamiński and V.V. Lozin, *Coloring edges and vertices of graphs without short or long cycles*, Contrib. Discrete. Math. **2** (2007), 61–66.

[14] R. M. Karp, *Reducibility Among Combinatorial Problems*, Complexity of Computer Computations, New York: Plenum., 85–103.

[15] D. Leven and Z. Galil, *NP-completeness of finding the chromatic index of regular graphs*, J. Algorithms **4** (1983), 35–44.

[16] B. Randerath and I. Schiermeyer, 3-*Colorability* $\in$ P *for* P_6*-free graphs*, Discrete Appl. Math. **136** (2004), 299–313.

[17] D. Seinsche, *On a property of the class of* n*-colorable graphs*, J. Comb. Theory B **16** (1974), 191–193.

[18] L. Stockmeyer, *Planar 3-colorability is polynomial complete*, SIGACT News (1973), 19–25.

[19] R.E. Tarjan, *Decomposition by clique separators*, Discrete Math. **55** (1985), 221–232.

Department of IEOR and Department of Mathematics, COLUMBIA UNIVERSITY, NEW YORK
E-mail: mchudnov@columbia.edu

Combinatorial theorems relative to a random set

David Conlon

Abstract. We describe recent advances in the study of random analogues of combinatorial theorems.

Mathematics Subject Classification (2010). Primary 05C80; Secondary 05C35.

Keywords. Extremal combinatorics, random graphs.

1. Introduction

Random graphs have played an integral role in extremal combinatorics since they were first used by Erdős [30] to prove an exponential lower bound for Ramsey numbers. The binomial random graph $G_{n,p}$ is a graph on n vertices where each of the $\binom{n}{2}$ possible edges is chosen independently with probability p. In modern terminology, Erdős' result says that with high probability $G_{n,1/2}$ contains no clique or independent set of order $2\log_2 n$. This then translates to a lower bound of $2^{t/2}$ for the Ramsey number $R(t)$ (this will be defined in Section 2).

Although there were several applications of random graphs prior to their work, the first systematic study of random graphs was undertaken by Erdős and Rényi [32, 33]. The concept of random graphs was also introduced independently by several other authors but, as explained by Bollobás [10], 'the other authors were all concerned with enumeration problems and their techniques were essentially deterministic.' Though it has its origins in applications to extremal combinatorics, the theory of random graphs is now a rich and self-sustaining area of study (see, for example, [10, 67]).

Suppose that $\mathcal{P}$ is a graph property, that is, a family of graphs closed under isomorphism. In studying random graphs, we are usually concerned with determining the probability that $G_{n,p}$ is in $\mathcal{P}$ for some property $\mathcal{P}$. For many properties, this probability exhibits a phase transition as p increases, changing abruptly from 0 to 1. The crossover point is known as the threshold. Formally, we say that $p^* := p^*(n)$ is a threshold for $\mathcal{P}$ if

$$\lim_{n\to\infty} \mathbb{P}[G_{n,p} \text{ is in } \mathcal{P}] = \begin{cases} 0 & \text{if } p = o(p^*), \\ 1 & \text{if } p = \omega(p^*). \end{cases}$$

We note that, depending on the property $\mathcal{P}$, the probability could also collapse from 1 to 0 as p increases. However, for most properties considered in this paper, the behaviour is as above. To give some simple examples, the properties of being connected and having a Hamiltonian cycle are both known to have a threshold at $p^* = \frac{\log n}{n}$, while the property of containing a

Proceedings of the International Congress of Mathematicians, Seoul, 2014

particular graph H has a threshold at $n^{-1/m(H)}$, where

$$m(H) = \max\left\{\frac{e(H')}{v(H')} : H' \subseteq H\right\}.$$

This function reflects the fact that a graph appears once its densest subgraph does.

Since the late eighties, there has been a great deal of interest in determining thresholds for analogues of combinatorial theorems to hold in random graphs and random subsets of other sets such as the integers. To give an example, we say that a graph G is K_3-Ramsey if any 2-colouring of the edges of G contains a monochromatic triangle. One of the foundational results in this area, proved by Frankl and Rödl [40] and Łuczak, Ruciński and Voigt [86], then states that there exists $C > 0$ such that if $p > C/\sqrt{n}$ then

$$\lim_{n\to\infty} \mathbb{P}[G_{n,p} \text{ is } K_3\text{-Ramsey}] = 1.$$

Frankl and Rödl used this theorem to prove that there are K_4-free graphs which are K_3-Ramsey, a result originally due to Folkman [38]. However, this new method allowed one to prove reasonable bounds for the size of such graphs, something which was not possible with previous methods.

From this beginning, a large number of papers were written on sparse random analogues of combinatorial theorems. These included papers on analogues of Ramsey's theorem, Turán's theorem and Szemerédi's theorem, though in many cases these efforts met with only partial success. This situation has changed dramatically in recent years and there are now three distinct, general methods for proving sparse random analogues of combinatorial theorems, furnishing solutions for many of the outstanding problems in the area.

The first two of these methods were developed by Gowers and the author [24] and, independently, by Schacht [112] and Friedgut, Rödl and Schacht [45]. The third method was found later by Balogh, Morris and Samotij [6] and, independently, by Saxton and Thomason [111]. Broadly speaking, the method employed by Gowers and the author builds on the transference principle developed by Green and Tao [60] in their proof that the primes contain arbitrarily long arithmetic progressions; the method of Schacht and Friedgut, Rödl and Schacht extends a multi-round exposure technique used by Rödl and Ruciński [98] in their study of Ramsey's theorem in random graphs; and the third method is a byproduct of general results about the structure of independent sets in hypergraphs, themselves building on methods of Kleitman and Winston [69] and Sapozhenko [108–110]. Of course, this summary does a disservice to all three methods, each of which involves the introduction of several new ideas. Surprisingly, all three proofs are substantially different and all three methods have their own particular strengths, some of which we will highlight below.

Rather than focusing on these three methods from the outset, we will further describe the developments leading up to them, explaining how these new results fit into the broader context. This will also allow us to review many of the important subsequent developments. We begin by discussing random analogues of Ramsey-type theorems.

2. Ramsey-type theorems in random sets

Ramsey's theorem [93] states that for any graph H and any natural number r there exists n such that any r-colouring of the edges of the complete graph K_n on n vertices contains a

monochromatic copy of H. The smallest such n is known as the r-colour Ramsey number of H and denoted $R(H;r)$. When $r=2$, we simply write this as $R(H)$ and when $H=K_t$, we just write $R(t)$. The result of Erdős mentioned in the introduction then says that $R(t)\geq 2^{t/2}$, while an upper bound due to Erdős and Szekeres [37] says that $R(t)\leq 4^t$. Though there have been lower order improvements to both of these estimates [17, 115], it remains a major open problem to give an exponential improvement to either of them.

Given a graph H and a natural number r, we say that a graph G is (H,r)-Ramsey if in any r-colouring of the edges of G there is guaranteed to be a monochromatic copy of H. Ramsey's theorem is the statement that K_n is (H,r)-Ramsey for n sufficiently large, while the overall aim of graph Ramsey theory is to decide which graphs are (H,r)-Ramsey for a given H and r. Though coNP-hard in general [15], this problem has borne much fruit and there is now a large theory with many interesting and important results (see, for example, [59]). One of the highlights of this theory is the following random Ramsey theorem of Rödl and Ruciński [96–98], which determines the threshold for Ramsey's theorem to hold in random graphs. As mentioned in the introduction, this result built on earlier work of Frankl and Rödl [40] and Łuczak, Ruciński and Voigt [86]. Here and throughout the paper, we will write $v(H)$ and $e(H)$ for the number of vertices and edges, respectively, of a graph H.

Theorem 2.1. *For any graph H that is not a forest consisting of stars and paths of length 3 and any positive integer $r\geq 2$, there exist positive constants c and C such that*

$$\lim_{n\to\infty}\mathbb{P}[G_{n,p}\text{ is }(H,r)\text{-Ramsey}]=\begin{cases}0 & \text{if } p<cn^{-1/m_2(H)},\\ 1 & \text{if } p>Cn^{-1/m_2(H)},\end{cases}$$

where

$$m_2(H)=\max\left\{\frac{e(H')-1}{v(H')-2}:H'\subseteq H \text{ and } v(H')\geq 3\right\}.$$

There are two parts to this theorem, one part saying that for $p<cn^{-1/m_2(H)}$ the random graph $G_{n,p}$ is highly unlikely to be (H,r)-Ramsey and the other saying that for $p>Cn^{-1/m_2(H)}$ it is almost surely (H,r)-Ramsey. Following standard usage, we will refer to these two parts as the 0-statement and the 1-statement, respectively.

The threshold in Theorem 2.1 occurs at $p^*=n^{-1/m_2(H)}$. This is the largest probability for which there is some subgraph H' of H such that the number of copies of H' in $G_{n,p}$ is approximately the same as the number of edges. For p significantly smaller than p^*, the number of copies of H' will also be significantly smaller than the number of edges. A rather delicate argument [96] then allows one to show that the edges of the graph may be colored so as to avoid any monochromatic copies of H'. For p significantly larger than p^*, almost every edge in the random graph is contained in many copies of every subgraph of H. The intuition, which takes substantial effort to make rigorous [98], is that these overlaps are enough to force the graph to be Ramsey.

That the proof of the 0-statement is delicate is betrayed by the omitted cases, which have smaller thresholds. For example, if a graph contains the star $K_{1,r(t-1)+1}$, then any r-colouring of the edges of this graph will contain a monochromatic $K_{1,t}$. However, the threshold for the appearance of $K_{1,r(t-1)+1}$ is lower than the threshold suggested by $m_2(K_{1,t})$. A more subtle case is when $H=P_4$, the path with 3 edges (and 4 vertices), and $r=2$. In this case, a cycle of length five with a pendant edge at each vertex is $(P_4,2)$-Ramsey. While the threshold for the appearance of these graphs is at n^{-1}, which is the same as $n^{-1/m_2(H)}$,

the threshold is coarse. This means that they start to appear with positive probability already when $p = c/n$ for any positive c. This implies that the 0-statement only holds when $p = o(1/n)$.

It is worth saying a little about the proof of the 1-statement in Theorem 2.1. We will focus on the case when $H = K_3$ and $r = 2$. The key idea is to write $G_{n,p}$ as the union of two independent random graphs G_{n,p_1} and G_{n,p_2}, chosen so that

$$p = p_1 + p_2 - p_1p_2 \quad \text{and} \quad p_2 = Lp_1$$

for some large constant L. We first expose the smaller random graph G_{n,p_1}. With high probability, every colouring of G_{n,p_1} will contain many monochromatic paths of length 2. If p_1 is a sufficiently large multiple of $1/\sqrt{n}$, it is also possible to show that with high probability these monochromatic paths are well distributed. In particular, for any given colouring of G_{n,p_1}, there are at least cn^3 triangles in the underlying graph K_n such that there is a path of the same colour, say red, between each pair of vertices in each triangle.

We now expose G_{n,p_2}. If this graph contains any of the cn^3 triangles described above, we are done, since each edge of this triangle must take the colour blue. Otherwise, together with the red connecting path, we would have a red triangle. By Janson's inequality [66], the probability that G_{n,p_2} does not contain any of the cn^3 triangles associated to this particular colouring is at most $2^{-c'p_2n^2}$, where c' depends on c. However, we must remember to account for every possible colouring of G_{n,p_1}. To do this, we take a union bound. Indeed, since there are at most $2^{p_1n^2}$ colourings of G_{n,p_1}, the probability that there exists a colouring such that G_{n,p_2} does not intersect the associated set of triangles is at most $2^{p_1n^2}2^{-c'p_2n^2}$. If we choose L sufficiently large, this probability tends to zero, completing the proof.

This method also allowed Rödl and Ruciński to determine the threshold for van der Waerden's theorem to hold in random subsets of the integers. Van der Waerden's theorem [124] states that for any natural numbers k and r there exists n such that any r-colouring of $[n] := \{1, 2, \dots, n\}$ contains a monochromatic k-term arithmetic progression, that is, a monochromatic subset of the form $\{a, a+d, \dots, a+(k-1)d\}$. To state the random version of this theorem, we define $[n]_p$ to be a random subset of $[n]$ where each element is chosen independently with probability p. We also say that a subset I of the integers is (k, r)-vdW if in any r-colouring of the points of I there is a monochromatic k-term arithmetic progression. Rödl and Ruciński's random van der Waerden theorem [98, 99] is then as follows.

Theorem 2.2. *For any positive integers $k \geq 3$ and $r \geq 2$, there exist positive constants c and C such that*

$$\lim_{n\to\infty} \mathbb{P}[[n]_p \textit{ is } (k,r)\textit{-vdW}] = \begin{cases} 0 & \text{if } p < cn^{-1/(k-1)}, \\ 1 & \text{if } p > Cn^{-1/(k-1)}. \end{cases}$$

The threshold is again a natural one, since it is the point where we expect that most vertices in $[n]_p$ will be contained in a constant number of k-term arithmetic progressions. We will say more about this in the next section when we discuss density theorems.

One question left open by the work of Rödl and Ruciński was whether Theorem 2.1 could be extended to hypergraphs. While some partial progress was made [100, 101], the general problem remained open, not least because of the apparent need to apply a hypergraph analogue of the regularity lemma, something which has only been developed in recent years [55, 88, 104, 120]. Approaches which circumvent hypergraph regularity were developed

independently by Friedgut, Rödl and Schacht [45] and by Gowers and the author [24], so that the following generalisation of Theorem 2.1 is now known. We write $G^{(k)}_{n,p}$ for the random k-uniform hypergraph on n vertices, where each edge is chosen independently with probability p.

Theorem 2.3. *For any k-uniform hypergraph H and any positive integer $r \geq 2$, there exists $C > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}[G^{(k)}_{n,p} \text{ is } (H,r)\text{-Ramsey}] = 1 \text{ if } p > Cn^{-1/m_k(H)},$$

where

$$m_k(H) = \max\left\{ \frac{e(H')-1}{v(H')-k} : H' \subseteq H \text{ and } v(H') \geq k+1 \right\}.$$

We note that the approach in [24] applies when H is strictly k-balanced, that is, when $m_k(H) > m_k(H')$ for every subgraph H' of H. However, almost all hypergraphs, including the complete hypergraph $K^{(k)}_t$, satisfy this requirement. A similar caveat applies to many of the theorems stated in this survey. We will usually make this explicit.

The 0-statement corresponding to Theorem 2.3 was considered by Gugelmann, Person, Steger and Thomas (see [61, 62]). In particular, their results imply the corresponding 0-statement for complete hypergraphs. However, there are again cases where the true threshold is smaller than $n^{-1/m_k(H)}$. Indeed, the picture seems to be more complicated than for graphs since there are examples other than the natural generalisations of paths and stars for which the 1-statement may be improved. We refer the reader to [61] for a more complete discussion.

One may also consider the threshold for asymmetric Ramsey properties. We say that a graph G is $(H_1, H_2, \ldots, H_r)$-Ramsey if any colouring of the edges of G with colours $1, 2, \ldots, r$ contains a monochromatic copy of H_i in colour i for some $i \in \{1, 2, \ldots, r\}$. A conjecture of Kohayakawa and Kreuter [71], which generalises Theorem 2.1, says that if $H_1, H_2, \ldots, H_r$ are graphs with $1 < m_2(H_r) \leq \cdots \leq m_2(H_1)$, then the $(H_1, H_2, \ldots, H_r)$-Ramsey property has a threshold at $n^{-1/m_2(H_1,H_2)}$, where

$$m_2(H_1,H_2) = \max\left\{ \frac{e(H_1')}{v(H_1') - 2 + 1/m_2(H_2)} : H_1' \subseteq H_1 \text{ and } v(H_1') \geq 3 \right\}.$$

Since the 0-statement fails to hold for certain forests, this statement should be qualified further, but it seems likely to hold for most collections of graphs.

Kohayakawa and Kreuter established the conjecture when $H_1, H_2, \ldots, H_r$ are cycles. As noted in [87], the same method shows that the KŁR conjecture (which we discuss in Section 4) would imply the 1-statement of the conjecture when H_1 is strictly 2-balanced, that is, when $m_2(H_1) > m_2(H_1')$ for all proper subgraphs H_1'. Since the KŁR conjecture is now an established fact, the following theorem is known to hold (as was noted explicitly by Balogh, Morris and Samotij [6]).

Theorem 2.4. *For any graphs $H_1, H_2, \ldots, H_r$ with $1 < m_2(H_r) \leq \cdots \leq m_2(H_1)$ and such that H_1 is strictly 2-balanced, there exists $C > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}[G_{n,p} \text{ is } (H_1, H_2, \ldots, H_r)\text{-Ramsey}] = 1 \text{ if } p > Cn^{-1/m_2(H_1,H_2)}.$$

A slightly weaker statement was established by Kohayakawa, Schacht and Spöhel [79] without appealing to the KŁR conjecture. Their proof is much closer in spirit to Rödl and

Ruciński's proof of Theorem 2.1. A corresponding 0-statement when $H_1, H_2, \ldots, H_r$ are cliques was established by Marciniszyn, Skokan, Spöhel and Steger [87]. However, the 0-statement remains open in general.

The methods developed in [24] and [45] also allow one to extend Rödl and Ruciński's results on random analogues of van der Waerden's theorem to a more general setting. A classical theorem of Rado [92] generalises van der Waerden's theorem by establishing necessary and sufficient conditions for a system of homogeneous linear equations

$$\sum_{j=1}^{k} a_{ij}x_j = 0 \text{ for } 1 \leq i \leq \ell$$

to be partition regular, that is, to be such that any finite colouring of the natural numbers contains a monochromatic solution $(x_1, x_2, \ldots, x_k)$ to this system of equations. To give an example, the solutions to the system of equations $x_i + x_{i+2} = 2x_{i+1}$ for $i = 1, 2, \ldots, k-2$ are k-term arithmetic progressions and so van der Waerden's theorem implies that this system of equations is partition regular. An extension of Theorem 2.2 was proved by Rödl and Ruciński in [99], but their 1-statement only applied to density regular systems of equations (though see also [57]). These are systems of equations, like the system defining k-term arithmetic progressions, whose solutions sets are closed under translation and dilation.

An extension of this theorem which applies to all partition regular systems of equations was proved by Friedgut, Rödl and Schacht [45]. More precisely, they proved a 1-statement, while the 0-statement had been established earlier by Rödl and Ruciński [99]. Since the details are somewhat technical, we refer the interested reader to [45] for further particulars.

We have already mentioned that the result of Frankl and Rödl [40] may be used to prove that there are K_4-free graphs which are $(K_3, 2)$-Ramsey. This was originally proved by Folkman [38] using a constructive argument. More generally, he proved that for any positive integer t there is a K_{t+1}-free graph which is $(K_t, 2)$-Ramsey. This beautiful result was subsequently extended to r-colourings by Nešetřil and Rödl [90, 91].

Once we know that these graphs exist, it is natural to try and estimate their size. We define the Folkman number $F(t)$ to be the smallest natural number n such that there exists a K_{t+1}-free graph G on n vertices with the property that every 2-colouring of the edges of G contains a monochromatic K_t. The upper bounds on $F(t)$ which come from the constructive proofs tend to have a dependency on t which, with a conservative estimate, is at least tower-type, that is, a tower of twos of height at least t. On the other hand, the lower bound is essentially the same as for Ramsey's theorem, that is, $F(t) \geq 2^{c't}$.

Very recently, it was noted that some of the methods for proving Ramsey-type theorems in random sets yield significantly stronger bounds for Folkman numbers [25, 102]. In particular, the following result was proved by Rödl, Ruciński and Schacht [102]. Their proof relies heavily on the hypergraph container results developed by Balogh, Morris and Samotij [6] and Saxton and Thomason [111] and an observation of Nenadov and Steger [89] that allows one to apply this machinery in the Ramsey setting.

Theorem 2.5. *There exists a constant c such that*

$$F(t) \leq 2^{ct^4 \log t}.$$

This bound is tantalisingly close to the lower bound and it would be of great interest to improve it further. Since we have now brought our discussions of Ramsey-type theorems in

random sets full circle, this provides a convenient departure point to move on to discussing density theorems in random sets, a topic about which much less was known before recent developments.

3. Density theorems in random sets

Turán's theorem [123] states that the largest K_t-free subgraph of K_n has at most $\left(1-\frac{1}{t-1}\right)\frac{n^2}{2}$ edges. Moreover, the unique K_t-free subgraph achieving this maximum is the $(t-1)$-partite graph with vertex sets $V_1, V_2, \dots, V_{t-1}$, where each set is of order $\lfloor \frac{n}{t-1} \rfloor$ or $\lceil \frac{n}{t-1} \rceil$. In particular, for $t = 3$, the triangle-free subgraph of K_n with the most edges is a bipartite graph with parts of order $\lfloor \frac{n}{2} \rfloor$ or $\lceil \frac{n}{2} \rceil$. A substantial generalisation of this theorem, known as the Erdős–Stone–Simonovits theorem [34, 36], states that for any graph H the largest H-free subgraph of K_n has at most $\left(1 - \frac{1}{\chi(H)-1} + o(1)\right)\binom{n}{2}$ edges, where $\chi(H)$ is the chromatic number of H.

We say that a graph G is (H, ϵ)-Turán if every subgraph of G with at least

$$\left(1 - \frac{1}{\chi(H)-1} + \epsilon\right) e(G)$$

edges contains a copy of H. The original work of Frankl and Rödl [40] on Ramsey properties in random graphs was actually motivated by a problem of Erdős and Nešetřil concerning an analogue of Folkman's theorem for the (H, ϵ)-Turán property. Specifically, they asked whether there exist K_4-free graphs which are (K_3, ϵ)-Turán and Frankl and Rödl showed that there are. Though not stated explicitly in their paper, Frankl and Rödl's method implies that for any $\epsilon > 0$ there exists $C > 0$ such that if $p > C/\sqrt{n}$ then

$$\lim_{n\to\infty} \mathbb{P}[G_{n,p} \text{ is } (K_3, \epsilon)\text{-Turán}] = 1.$$

Unlike Ramsey properties, the corresponding 0-statement is easy to prove. Indeed, for p a sufficiently small multiple of $1/\sqrt{n}$, the number of triangles in $G_{n,p}$ will be significantly smaller than the number of edges. We may therefore remove all copies of K_3 by deleting one edge from each copy, leaving a subgraph which is triangle-free but contains at least $(1-\delta)e(G_{n,p})$ edges.

A similar argument provides a lower bound for all H. That is, if the number of copies of H is significantly smaller than the number of edges, we can remove all copies of H by deleting one edge from each copy. Therefore, if $p^{e(H)}n^{v(H)} \ll pn^2$, that is, $p \ll n^{-(v(H)-2)/(e(H)-1)}$, the (H, ϵ)-Turán property cannot hold. Since the same argument applies for any subgraph H' of H, it is easy to see that for $p \ll n^{-1/m_2(H)}$ the random graph $G_{n,p}$ cannot be (H, ϵ)-Turán. Here $m_2(H)$ is defined as in Theorem 2.1, that is,

$$m_2(H) = \max\left\{\frac{e(H')-1}{v(H')-2} : H' \subseteq H \text{ and } v(H') \geq 3\right\},$$

The natural conjecture that the (H, ϵ)-Turán property holds in random graphs with $p \gg n^{-1/m_2(H)}$ was first stated by Haxell, Kohayakawa and Łuczak [63, 64] and reiterated by Kohayakawa, Łuczak and Rödl [73].

Until recently, this conjecture was only known to hold for a small collection of graphs, including K_3, K_4 and K_5 [40, 53, 73] and all cycles [46, 63, 64] (see also [76, 117]). A verification of the conjecture for all graphs was completed by Schacht [112] and by Gowers and the author [24], although we must qualify this statement by saying that the results of [24] apply when H is strictly 2-balanced, that is, when $m_2(H') < m_2(H)$ for all $H' \subset H$. However, the class of strictly 2-balanced graphs includes many of the graphs one normally considers, such as cliques and cycles.

Theorem 3.1. *For any graph H and any $\epsilon > 0$, there exist positive constants c and C such that*

$$\lim_{n\to\infty} \mathbb{P}[G_{n,p} \text{ is } (H,\epsilon)\text{-Turán}] = \begin{cases} 0 & \text{if } p < cn^{-1/m_2(H)}, \\ 1 & \text{if } p > Cn^{-1/m_2(H)}. \end{cases}$$

As mentioned in the introduction, Schacht's proof of Theorem 3.1 builds on Rödl and Ruciński's proof of Theorem 2.1. In the last section, we gave a brief description of their method, showing how it was best to think of the random graph $G_{n,p}$ as the union of two independent random graphs G_{n,p_1} and G_{n,p_2}. In Schacht's method, this multi-round exposure is taken further, the rough idea being to expose $G_{n,p}$ over several successive rounds and to apply a density increment argument.

The method employed in [24] relies upon proving a transference principle, an idea which originates in the work of Green and Tao [60] (see also [56, 94]). In the case of triangles, this transference principle says that for $p \geq C/\sqrt{n}$ any subgraph G of $G_{n,p}$ may be modelled by a subgraph K of the complete graph K_n in such a way that the proportion of edges and triangles in K is close to the proportion of edges and triangles in G. That is, if the sparse graph G contains $c_1 pn^2$ edges and $c_2 p^3 n^3$ triangles, then the dense model K will contain approximately $c_1 n^2$ edges and $c_2 n^3$ triangles.

Suppose now that we wish to prove Turán's theorem for triangles relative to a random graph. Given a subgraph G of $G_{n,p}$ with $\left(\frac{1}{2}+\epsilon\right) p\binom{n}{2}$ edges, we know, once our approximation is sufficiently good, that its dense model K has at least $\left(\frac{1}{2}+\frac{\epsilon}{2}\right)\binom{n}{2}$ edges. A robust version of Turán's theorem [35] then implies that K contains at least cn^3 triangles for some $c > 0$ depending on ϵ. Provided again that our approximation is sufficiently good, this implies that G contains at least $\frac{c}{2}p^3n^3$ triangles, which is even more than we required.

Though the analogue of Turán's theorem for hypergraphs is rather poorly understood (see, for example, [68]), a similar strategy shows that it is still possible to transfer it to the random setting. To state the result, we need some definitions. Given a k-uniform hypergraph H, we let $\mathrm{ex}(n,H)$ be the largest number of edges in an H-free subgraph of $K_n^{(k)}$ and

$$\pi_k(H) = \lim_{n\to\infty} \frac{\mathrm{ex}(n,H)}{\binom{n}{k}}.$$

We then say that a k-uniform hypergraph G is (H,ϵ)-Turán if every subgraph of G with at least $(\pi_k(H)+\epsilon)\, e(G)$ edges contains a copy of H. Let $m_k(H)$ be defined as in the previous subsection, that is,

$$m_k(H) = \max\left\{ \frac{e(H')-1}{v(H')-k} : H' \subseteq H \text{ and } v(H') \geq k+1 \right\}.$$

Then the analogue of Theorem 3.1, proved in [24, 112], states that the property of being (H,ϵ)-Turán for a k-uniform hypergraph H has a threshold at $n^{-1/m_k(H)}$.

Theorem 3.2. *For any k-uniform hypergraph H and any $\epsilon > 0$, there exist positive constants c and C such that*

$$\lim_{n\to\infty} \mathbb{P}[G^{(k)}_{n,p} \text{ is } (H,\epsilon)\text{-Turán}] = \begin{cases} 0 & \text{if } p < cn^{-1/m_k(H)}, \\ 1 & \text{if } p > Cn^{-1/m_k(H)}. \end{cases}$$

One structural counterpart to Turán's theorem is the Erdős-Simonovits stability theorem [114]. This says that for any graph H with $\chi(H) \geq 3$ and any $\epsilon > 0$, there exists $\delta > 0$ such that any H-free subgraph of K_n with at least $\left(1 - \frac{1}{\chi(H)-1} - \delta\right)\binom{n}{2}$ edges may be made $(\chi(H)-1)$-partite by removing at most ϵn^2 edges. The following sparse analogue of this result was originally proved in [24] for strictly 2-balanced graphs. Later, Samotij [107] found a way to amend Schacht's method so that it applied to stability statements, extending this result to all graphs.

Theorem 3.3. *For any graph H with $\chi(H) \geq 3$ and any $\epsilon > 0$, there exist positive constants δ and C such that if $p \geq Cn^{-1/m_2(H)}$ the random graph $G_{n,p}$ a.a.s. has the following property. Every H-free subgraph of $G_{n,p}$ with at least $\left(1 - \frac{1}{\chi(H)-1} - \delta\right)p\binom{n}{2}$ edges can be made $(\chi(H)-1)$-partite by removing at most ϵpn^2 edges.*

For cliques, Turán's theorem has a much more precise corresponding structural statement, saying that the largest K_t-free subgraph is $(t-1)$-partite. One may therefore ask when this property holds a.a.s. in the random graph $G_{n,p}$. This question was first studied by Babai, Simonovits and Spencer [4] who showed that for $p > \frac{1}{2}$ the size of the maximum triangle-free subgraph is a.a.s. the same as the size of the largest bipartite subgraph. This result was extended to the range $p > n^{-c}$ by Brightwell, Panagiotou and Steger [14]. Recently, DeMarco and Kahn [27] proved the following much more precise result.

Theorem 3.4. *There is a positive constant C such that if $p > C\sqrt{\log n/n}$ then a.a.s. every maximum triangle-free subgraph of $G_{n,p}$ is bipartite.*

The threshold here is different from the $1/\sqrt{n}$ we have come to expect. However, the result is sharp up to the constant C. Indeed, for $p = 0.1\sqrt{\log n/n}$, the random graph $G_{n,p}$ will typically contain a 5-cycle none of whose edges are contained in a triangle. In a forthcoming paper, DeMarco and Kahn [28] prove the following extension of this result to all cliques. Once again, the extra log factors are essential.

Theorem 3.5. *For any natural number t, there exists $C > 0$ such that if*

$$p > Cn^{-\frac{2}{t+1}} \log^{\frac{2}{(t+1)(t-2)}} n$$

then a.a.s. every maximum K_t-free subgraph of $G_{n,p}$ is $(t-1)$-partite.

We note that a related question, where one wishes to determine the range of m for which most K_t-free graphs with n vertices and m edges are $(t-1)$-partite, was solved recently by Balogh, Morris, Samotij and Warnke [7].

The methods of [24] and [112] also allow one to prove sparse analogues of density statements from other settings. For example, Szemerédi's theorem [118] states that for any natural number k and any $\delta > 0$ there exists n_0 such that if $n \geq n_0$ any subset of $[n]$ of density at least δ contains a k-term arithmetic progression. This is the density version of van der

Waerden's theorem and trivially implies that theorem by taking $\delta = \frac{1}{r}$ and considering the largest colour class. This theorem and the tools arising in its many proofs [48, 54, 95] have been enormously influential in the development of modern combinatorics.

We say that a subset I of the integers is (k, δ)-Szemerédi if any subset of I with at least $\delta|I|$ elements contains an arithmetic progression of length k. Szemerédi's theorem says that for n sufficiently large the set $[n]$ is (k, δ)-Szemerédi, while a striking corollary of Green and Tao's work on arithmetic progressions in the primes [60] says that for n sufficiently large the set of primes up to n is (k, δ)-Szemerédi.

For random subsets of the integers, the (k, δ)-Szemerédi property was first studied by Kohayakawa, Łuczak and Rödl [72], who proved that the property of being $(3, \delta)$-Szemerédi has a threshold at $1/\sqrt{n}$. In general, the natural conjecture is that the (k, δ)-Szemerédi property has a threshold at $n^{-1/(k-1)}$. The lower bound is again straightforward, since for $p \ll n^{-1/(k-1)}$ the number of k-term arithmetic progressions is significantly smaller than the number of elements in the random set $[n]_p$, allowing us to remove one element from each arithmetic progression without significantly affecting the density. The corresponding 1-statement was provided in [24] and [112].

Theorem 3.6. *For any integer $k \geq 3$ and $\delta > 0$, there exist positive constants c and C such that*

$$\lim_{n\to\infty} \mathbb{P}[[n]_p \textit{ is } (k,\delta)\textit{-Szemerédi}] = \begin{cases} 0 & \textit{if } p < cn^{-1/(k-1)}, \\ 1 & \textit{if } p > Cn^{-1/(k-1)}. \end{cases}$$

A particularly satisfying approach to density theorems in random sets is provided by the recent hypergraph containers method of Balogh, Morris and Samotij [6] and Saxton and Thomason [111], the only probabilistic input being Chernoff's inequality and the union bound. In the context of Szemerédi's theorem, one of the main corollaries of this method is the following theorem.

Theorem 3.7. *For any integer $k \geq 3$ and any $\epsilon > 0$, there exists $C > 0$ such that if $m \geq Cn^{1-1/(k-1)}$, then there are at most $\binom{\epsilon n}{m}$ subsets of $\{1, 2, \ldots, n\}$ of order m which contain no k-term arithmetic progression.*

Given this statement, which is completely deterministic, it is straightforward to derive the 1-statement in Theorem 3.6, so much so that we may now give the entire calculation. For brevity, we write (k, δ)-Sz rather than (k, δ)-Szemerédi and $\mathcal{I}_k(n, \delta pn/2)$ for the collection of subsets of $\{1, 2, \ldots, n\}$ of order $\delta pn/2$ which contain no k-term arithmetic progression. We have

$$\begin{aligned}
\mathbb{P}[[n]_p \text{ is not } (k,\delta)\text{-Sz}] &\leq \mathbb{P}[|[n]_p| < pn/2] + \mathbb{P}[|[n]_p| \geq pn/2 \text{ and } [n]_p \text{ is not } (k,\delta)\text{-Sz}] \\
&\leq \exp(-\Omega(pn)) + \mathbb{P}[[n]_p \supseteq I \text{ for some } I \in \mathcal{I}_k(n, \delta pn/2)] \\
&\leq \exp(-\Omega(pn)) + \binom{\epsilon n}{\delta pn/2} p^{\delta pn/2} \\
&\leq \exp(-\Omega(pn)) + \left(\frac{2e\epsilon pn}{\delta pn}\right)^{\delta pn/2} \\
&= \exp(-\Omega(pn)),
\end{aligned}$$

provided $\epsilon < \delta/2e$.

Deriving Theorem 3.1 from the results of [6] and [111] involves a little more work. To describe the idea, we focus on the case where $H = K_3$. We begin by considering the 3-uniform hypergraph $\mathcal{G}$ whose vertex set is the collection of edges in K_n and whose edge set is the collection of triangles in K_n. Turán's theorem for triangles may then be restated as saying that this 3-uniform hypergraph has no independent set of order greater than $\left(\frac{1}{2} + o(1)\right)|V(\mathcal{G})|$. We would now like to show that if $p \geq C/\sqrt{n}$ then the random set $V(\mathcal{G})_p$ formed by choosing each element of $V(\mathcal{G})$ independently with probability p contains no independent set of order greater than $\left(\frac{1}{2} + \epsilon\right) p|V(\mathcal{G})|$.

One approach would be to use the union bound and Chernoff's inequality to show that with high probability the intersection of the random set with each independent set is as required. An argument of this variety worked in the proof of Theorem 3.6 above, but usually there are far too many independent sets for this approach to be feasible. The main results in both [6] and [111] circumvent this difficulty by showing that there is a substantially smaller collection of almost independent sets which contain all independent sets. Since these sets are almost independent, we know, by the robust version of Turán's theorem, that they must also have size at most $\left(\frac{1}{2} + \frac{\epsilon}{2}\right)|V(\mathcal{G})|$, say. Applying the union bound over this smaller set then allows us to derive the result.

4. Regularity in random graphs

Szemerédi's regularity lemma [119] is one of the cornerstones of modern graph theory (see [81, 103]). Roughly speaking, it says that the vertex set of every graph G may be divided into a bounded number of parts in such a way that most of the induced bipartite graphs between different parts are pseudorandom. To be more precise, we need some definitions.

We say that a bipartite graph between sets U and V is ϵ-regular if, for every $U' \subseteq U$ and $V' \subseteq V$ with $|U'| \geq \epsilon|U|$ and $|V'| \geq \epsilon|V|$, the density $d(U', V')$ of edges between U' and V' satisfies

$$|d(U', V') - d(U, V)| \leq \epsilon.$$

A partition of the vertex set of a graph into t pieces $V_1, \ldots, V_t$ is an equipartition if, for every $1 \leq i, j \leq t$, we have $||V_i| - |V_j|| \leq 1$. Finally, a partition is ϵ-regular if it is an equipartition and, for all but at most ϵt^2 pairs (V_i, V_j), the induced graph between V_i and V_j is ϵ-regular. Szemerédi's regularity lemma can now be stated as follows.

Theorem 4.1. *For any $\epsilon > 0$, there exists an integer T such that every graph G admits an ϵ-regular partition $V_1, \ldots, V_t$ of its vertex set into $t \leq T$ pieces.*

For sparse graphs – that is, graphs with n vertices and $o(n^2)$ edges – the regularity lemma stated above is vacuous, since every equipartition into a bounded number of parts is ϵ-regular for n sufficiently large. However, as observed independently by Kohayakawa [70] and Rödl, there is a meaningful analogue of the regularity lemma for sparse graphs, provided one is willing to restrict consideration to a well-behaved class of graphs.

To make this more precise, we say that a bipartite graph between sets U and V is (ϵ, p)-regular if, for every $U' \subseteq U$ and $V' \subseteq V$ with $|U'| \geq \epsilon|U|$ and $|V'| \geq \epsilon|V|$, the density $d(U', V')$ of edges between U' and V' satisfies

$$|d(U', V') - d(U, V)| \leq \epsilon p.$$

That is, we alter the definition of regularity so that it is relative to a particular density p, usually chosen to be comparable to the total density between U and V. A partition of the vertex set of a graph into t pieces $V_1, \dots, V_t$ is then said to be (ϵ, p)-regular if it is an equipartition and, for all but at most ϵt^2 pairs (V_i, V_j), the induced graph between V_i and V_j is (ϵ, p)-regular.

The class of graphs to which the Kohayakawa-Rödl regularity lemma applies are the so-called upper-uniform graphs [75]. Suppose that $0 < \eta \leq 1$, $D > 1$ and $0 < p \leq 1$ are given. We will say that a graph G is (η, p, D)-upper-uniform if for all disjoint subsets U_1 and U_2 with $|U_1|, |U_2| \geq \eta |V(G)|$, the density of edges between U_1 and U_2 satisfies $d(U_1, U_2) \leq Dp$. This condition is satisfied for many natural classes of graphs, including all subgraphs of random graphs of density p. The sparse regularity lemma of Kohayakawa and Rödl is now as follows.

Theorem 4.2. *For any $\epsilon > 0$ and $D > 1$, there exists $\eta > 0$ and an integer T such that for every $p \in [0, 1]$, every graph G that is (η, p, D)-upper-uniform admits an (ϵ, p)-regular partition $V_1, \dots, V_t$ of its vertex set into $t \leq T$ pieces.*

A recent variant of this lemma, due to Scott [113], requires no upper-uniformity assumption on G, although it is often useful to impose such a constraint in practice. Since the two statements are interchangeable when one is dealing with a subgraph of the random graph, we have chosen to describe the original version.

In applications, the regularity method is usually applied in combination with a counting lemma. Roughly speaking, a counting lemma says that if we start with an arbitrary graph H and replace its vertices by large independent sets and its edges by ϵ-regular bipartite graphs with non-negligible density, then this blow-up will contain roughly the expected number of copies of H. To state this result formally, we again need some definitions.

Given a graph H with vertex set $\{1, 2, \dots, k\}$ and a collection of disjoint vertex sets $V_1, V_2, \dots, V_k$ in a graph G, we say that a k-tuple $(v_1, v_2, \dots, v_k)$ is a canonical copy of H in G if $v_i \in V_i$ for every $i \in V(H)$ and $v_i v_j \in E(G)$ for every $ij \in E(H)$. We write $G(H)$ for the number of canonical copies of H in G. The counting lemma may now be stated as follows.

Lemma 4.3. *For any graph H with vertex set $\{1, 2, \dots, k\}$ and any $\delta > 0$, there exists a positive constant ϵ and an integer n_0 such that the following holds. Let $n \geq n_0$ and let G be a graph whose vertex set is a disjoint union $V_1 \cup V_2 \cup \dots \cup V_k$ of sets of size n. Assume that for each $ij \in E(H)$, the bipartite subgraph of G induced by V_i and V_j is ϵ-regular and has density d_{ij}. Then*

$$G(H) = \left(\prod_{ij \in E(H)} d_{ij} \pm \delta \right) n^k.$$

When combined with the regularity lemma, this result allows one to prove a number of well-known theorems in extremal graph theory, including the Erdős–Stone–Simonovits theorem, its stability version and the graph removal lemma. In order to extend these results to sparse graphs, one plausible approach, championed by Kohayakawa, Łuczak and Rödl [73], would be to extend Lemma 4.3 to sparse graphs. For example, it would be ideal if we could replace the densities d_{ij} with $d_{ij}p$, the ϵ-regularity condition with an (ϵ, p)-regularity

condition and the conclusion with

$$G(H) = \left(\prod_{ij \in E(H)} d_{ij} \pm \delta \right) p^{e(H)} n^k.$$

We will initially aim for less, only asking to embed a single canonical copy of H. Unfortunately, for reasons with which we are now familiar, we cannot hope that such a statement holds for small p. Indeed, if $p \ll n^{-1/m_2(H)}$, there is a subgraph H' of H for which $p^{e(H')} n^{v(H')} \ll pn^2$. We may therefore remove all copies of H', and hence H, from $G_{n,p}$ while deleting only a small fraction of the edges. The resulting graph is both (ϵ, p)-regular, for some small ϵ, and H-free.

Frustratingly, this embedding lemma also fails for larger values of p. To see this, take a counterexample of the kind just described but with the sets V_i of order r for some r that is much smaller than n. Now replace each vertex of this small graph by an independent set with n/r vertices and each edge with a complete bipartite graph. This yields a graph with n vertices in each V_i. It is easy to see that the counterexample survives this blowing-up process, implying that the sought-after sparse embedding lemma is false whenever $p = o(1)$ (see [52, 74]).

However, these counterexamples have a very special structure, an observation that led Kohayakawa, Łuczak and Rödl to conjecture that they might be rare. Roughly speaking, their conjecture, known as the KŁR conjecture, stated that if $p \gg n^{-1/m_2(H)}$, then the number of counterexamples to the embedding lemma is so small that $G_{n,p}$ should not typically contain any such counterexample as a subgraph. Before stating the conjecture (or theorem as it is now), we introduce some notation.

As above, let H be a graph with vertex set $\{1, 2, \ldots, k\}$. We denote by $\mathcal{G}(H, n, m, p, \epsilon)$ the collection of all graphs G obtained in the following way. The vertex set of G is a disjoint union $V_1 \cup V_2 \cup \cdots \cup V_k$ of sets of size n. For each edge $ij \in E(H)$, we add an (ϵ, p)-regular bipartite graph with m edges between the pair (V_i, V_j). These are the only edges of G. We also write $\mathcal{G}^*(H, n, m, p, \epsilon)$ for the set of all $G \in \mathcal{G}(H, n, m, p, \epsilon)$ that do not contain a canonical copy of H.

Since the sparse regularity lemma could yield graphs with different densities between the various pairs of vertex sets, it may seem surprising that we are restricting attention to graphs where all the densities are equal. However, it is sufficient to consider just this case. In fact, the KŁR conjecture, which we now state, is more specific still, since it also takes $p = m/n^2$. Again, it turns out that from this case one can deduce any other cases that may be needed.

Theorem 4.4. *Let H be a fixed graph and let $\beta > 0$. Then there exist positive constants C and ϵ such that*

$$|\mathcal{G}^*(H, n, m, m/n^2, \epsilon)| \leq \beta^m \binom{n^2}{m}^{e(H)}$$

for every $m \geq Cn^{2-1/m_2(H)}$.

The KŁR conjecture has attracted considerable attention over the past two decades and was resolved for a number of special cases. The cases $H = K_3$, K_4 and K_5 were solved in [72], [51], and [53], respectively. For cycles, the conjecture was proved in [9, 49] (see also [71] for a slightly weaker version). Related results were also given in [50] and [76]. We state it as a theorem because it has now been proved in full generality by Balogh, Morris and Samotij [6] and by Saxton and Thomason [111].

Many of the results discussed in this survey, including Theorems 3.1 and 3.3, follow easily from the KŁR conjecture. Indeed, these applications were the original motivation for the conjecture. However, there are situations where an embedding result is not enough: rather than just a single copy of H, one needs to know that there are many copies. That is, one needs something more like a full counting lemma. Such a counting lemma was provided in a paper of Gowers, Samotij, Schacht and the author [26], the main result of which is the following. We allow for different densities between parts by replacing m with a vector $\mathbf{m} = (m_{ij})_{ij \in E(H)}$.

Theorem 4.5. *For any graph H and any $\delta, d > 0$, there exist positive constants ϵ and ξ with the following property. For any $\eta > 0$, there is $C > 0$ such that if $p \geq CN^{-1/m_2(H)}$ then a.a.s. the following holds in $G_{N,p}$:*

(i) *For any $n \geq \eta N$, $\mathbf{m}$ with $m_{ij} \geq dpn^2$ for all $ij \in E(H)$ and any subgraph G of $G_{N,p}$ in $\mathcal{G}(H, n, \mathbf{m}, p, \epsilon)$,*

$$G(H) \geq \xi \left(\prod_{ij \in E(H)} \frac{m_{ij}}{n^2} \right) n^{v(H)}.$$

(ii) *Moreover, if H is strictly 2-balanced, then*

$$G(H) = (1 \pm \delta) \left(\prod_{ij \in E(H)} \frac{m_{ij}}{n^2} \right) n^{v(H)}.$$

We note that Theorem 4.5(i) follows from Samotij's adaptation [107] of Schacht's method [112] (and may also be derived from the work of Saxton and Thomason [111]), while Theorem 4.5(ii) follows from the work of Gowers and the author [24]. Though stronger than Theorem 4.4 in some obvious ways, it is worth noting that Theorem 4.5 does not return the estimate for the number of counterexamples provided by that theorem. This estimate is important for some applications, Theorem 2.4 being a notable example.

One sample application where we need a counting result rather than an embedding result is for proving a random analogue of the graph removal lemma. This theorem, usually attributed to Ruzsa and Szemerédi [106] (though see also [3, 31, 47]), is as follows: for any $\delta > 0$, there exists $\epsilon > 0$ such that if G is a graph on n vertices containing at most $\epsilon n^{v(H)}$ copies of H, then G may be made H-free by deleting at most δn^2 edges. Though simple in appearance, this result is surprisingly difficult to prove (see, for example, [18, 39]). It also has some striking consequences, including the $k = 3$ case of Szemerédi's theorem, originally due to Roth [105]. A sparse random version of the graph removal lemma was conjectured by Łuczak in [85] and proved, for strictly 2-balanced H, in [24]. The following statement, which applies for all H, may be found in [26].

Theorem 4.6. *For any graph H and any $\delta > 0$, there exist positive constants ϵ and C such that if $p \geq Cn^{-1/m_2(H)}$ then the following holds a.a.s. in $G_{n,p}$. Every subgraph of $G_{n,p}$ which contains at most $\epsilon p^{e(H)} n^{v(H)}$ copies of H may be made H-free by removing at most δpn^2 edges.*

Note that if $p \leq cn^{-1/m_2(H)}$, for c sufficiently small, this statement is trivially true. Indeed, in this range, there exists a subgraph H' of H such that the number of copies of

H' in $G_{n,p}$ is smaller than δpn^2, so we can simply remove one edge from each copy of H'. One might then conjecture, as Łuczak did, that Theorem 4.6 holds for all values of p. For 2-balanced graphs, those with $m_2(H') \leq m_2(H)$ for all $H' \subset H$, we may verify this conjecture by taking ϵ to be sufficiently small in terms of C, δ, and H. For $p \leq Cn^{-1/m_2(H)}$ and $\epsilon < \delta C^{-e(H)}$, the number of copies of H is at most $\epsilon p^{e(H)} n^{v(H)} \leq \epsilon C^{e(H)} pn^2 < \delta pn^2$. Deleting an edge from each copy yields the result.

5. Further directions

5.1. Sharp thresholds for Ramsey properties. A graph property $\mathcal{P}$ is said to be monotone if it is closed under the addition of edges, that is, $G \in \mathcal{P}$ and $G \subset G'$ implies that $G' \in \mathcal{P}$. A result of Bollobás and Thomason [11] shows that any monotone property has a threshold. For example, since Ramsey properties are clearly monotone, this immediately implies that the (H, r)-Ramsey property and the (k, r)-vdW property, both defined in Section 2, have thresholds.

Once we have proved that a given property has a threshold, it is often interesting to study this threshold more closely. We say that $\mathcal{P}$ has a sharp threshold at $p^* := p^*(n)$ if, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[G_{n,p} \text{ is in } \mathcal{P}] = \begin{cases} 0 & \text{if } p < (1-\epsilon)p^*, \\ 1 & \text{if } p > (1+\epsilon)p^*. \end{cases}$$

For example, the properties of being connected and having a Hamiltonian cycle have sharp thresholds, while the property of containing a particular graph H has a non-sharp or coarse threshold.

A seminal result of Friedgut [41] gives a criterion for assessing whether a monotone property has a sharp threshold or not. Roughly speaking, this criterion says that if the property is globally determined the threshold is sharp, while if it is locally determined it is not. This fits in well with the examples given above, since connectedness and Hamiltonicity are clearly global properties, while having a single copy of a particular H is decidedly local.

The question of whether Ramsey properties have sharp thresholds was first studied by Friedgut and Krivelevich [43]. They proved, amongst other things, that the (H, r)-Ramsey property is sharp when H is any tree other than a star or a path of length three. However, the first substantial breakthrough was made by Friedgut, Rödl, Ruciński and Tetali [44], who proved that the $(K_3, 2)$-Ramsey property has a sharp threshold. Their result may be stated as follows.

Theorem 5.1. *There exists a bounded function $\hat{c} := \hat{c}(n)$ such that for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}[G_{n,p} \textit{ is } (K_3, 2)\textit{-Ramsey}] = \begin{cases} 0 & \textit{if } p < (1-\epsilon)\hat{c}/\sqrt{n}, \\ 1 & \textit{if } p > (1+\epsilon)\hat{c}/\sqrt{n}. \end{cases}$$

A close look at this result reveals an unusual feature: though we know that the threshold is sharp, we do not know exactly where it lies. In principle, the function $\hat{c}(n)$ could depend on n and wander up and down between constants c and C. However, we expect that the true behaviour should be that it tends towards a constant. It would be very interesting to prove that this is the case. It would also be of great interest to extend Theorem 5.1 to other graphs and a higher number of colours.

More recently, Friedgut, Hàn, Person and Schacht [42] proved that there is a sharp threshold for the appearance of k-term arithmetic progressions in every 2-colouring of $[n]_p$. That is, they showed that the $(k,2)$-vdW property has a sharp threshold. Their proof relies in a fundamental way on the hypergraph containers results discussed throughout this survey.

Theorem 5.2. *For every integer $k \geq 3$, there exists a bounded function $\hat{c}_k := \hat{c}_k(n)$ such that for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}[G_{n,p} \text{ is } (k,2)\text{-vdW}] = \begin{cases} 0 & \text{if } p < (1-\epsilon)\hat{c}_k n^{-1/(k-1)}, \\ 1 & \text{if } p > (1+\epsilon)\hat{c}_k n^{-1/(k-1)}. \end{cases}$$

It would again be interesting to determine the asymptotic behaviour of $\hat{c}_k(n)$ or to extend this result to a higher number of colours.

5.2. Large subgraph theorems in random graphs. One of the most active areas of research in extremal combinatorics is in finding conditions under which a graph contains certain large or even spanning sparse subgraphs (see, for example, [83]). It is therefore natural to ask whether these results also have random analogues.

One of the standard examples in this area is Dirac's theorem [29], which says that if a graph on n vertices has minimum degree at least $n/2$ then it contains a Hamiltonian cycle, that is, a cycle which meets every vertex. The study of random analogues of Dirac's theorem was initiated by Sudakov and Vu [116] and the state of the art is now the following result of Lee and Sudakov [84].

Theorem 5.3. *For any $\epsilon > 0$, there exists $C > 0$ such that if $p \geq C\frac{\log n}{n}$ then a.a.s. every subgraph of $G_{n,p}$ with minimum degree at least $\left(\frac{1}{2}+\epsilon\right)pn$ contains a Hamiltonian cycle.*

There has also been considerable work on studying random analogues of the bandwidth theorem of Böttcher, Schacht and Taraz [13]. The bandwidth of a graph G is the smallest b for which there is an ordering $v_1, v_2, \ldots, v_n$ of the vertices of G such that $|i-j| \leq b$ for all edges v_iv_j. The theorem then states that for any positive integers r and Δ and any $\gamma > 0$, there exists an integer n_0 and $\beta > 0$ such that if $n \geq n_0$ and H is an n-vertex graph with chromatic number r, maximum degree Δ and bandwidth at most βn, then any graph on n vertices with minimum degree at least $\left(1-\frac{1}{r}+\gamma\right)n$ contains a copy of H.

For the $r = 2$ case, that is, for bipartite H, the following random analogue of this theorem was proved by Böttcher, Kohayakawa and Taraz [12].

Theorem 5.4. *For any integer $\Delta \geq 2$ and any $\eta, \gamma > 0$, there exist positive constants β and C such that if $p \geq C(\log n/n)^{1/\Delta}$ the random graph $G_{n,p}$ a.a.s. has the following property. Any subgraph of $G_{n,p}$ with minimum degree at least $\left(\frac{1}{2}+\gamma\right)pn$ contains any bipartite graph on at most $(1-\eta)n$ vertices with maximum degree Δ and bandwidth at most βn.*

Related results were also proved by Huang, Lee and Sudakov [65]. In particular, they showed that if H is an r-partite graph on n vertices such that every vertex is contained in a triangle, then there exist subgraphs of the random graph $G_{n,p}$ with minimum degree at least $\left(1-\frac{1}{r}+\gamma\right)pn$ such that at least cp^{-2} vertices are not contained in a copy of H. That is, we cannot hope to cover all vertices when considering random analogues of the bandwidth theorem. However, as suggested by results in [5] and [65], it may still be possible to embed graphs with as many as $n - Cp^{-2}$ vertices.

A celebrated result of Chvátal, Rödl, Szemerédi and Trotter [16] (see also [19, 58]) states that for any positive integers Δ and r, there exists $C > 0$ such that if H is any graph with n vertices and maximum degree Δ, then $R(H;r) \leq Cn$. That is, the Ramsey number of bounded degree graphs grows linearly in the number of vertices. However, one can do even better.

Given a graph H and a natural number r, we define the size-Ramsey number $\hat{R}(H;r)$ to be the smallest number of edges in an (H,r)-Ramsey graph. So we are now interested in minimising the number of edges rather than the number of vertices. A striking result of Beck [8] says that $\hat{R}(P_n;r) \leq Cn$ for some C depending only on r. Using random graphs, Kohayakawa, Rödl, Schacht and Szemerédi [78] recently proved that if H is any graph with n vertices and maximum degree Δ, then $\hat{R}(H;r) \leq n^{2-\frac{1}{\Delta}+o(1)}$. That is, the size-Ramsey number of bounded degree graphs is subquadratic in the number of vertices. Precisely stated, their main result is the following.

Theorem 5.5. *For any integers $\Delta \geq 2$ and $r \geq 2$, there exists $C > 0$ such that if $p \geq C(\log N/N)^{1/\Delta}$ the random graph $G_{N,p}$ with $N = Cn$ a.a.s. has the following property. Any r-colouring of the edges of $G_{n,p}$ contains a colour class which contains every graph on n vertices with maximum degree Δ.*

In a forthcoming paper, Allen, Böttcher, Hàn, Kohayakawa and Person [1] prove a sparse random version of the blow-up lemma. For dense graphs, this result, proved by Komlós, Sárközy and Szemerédi [80], is a standard tool for embedding spanning subgraphs. Its sparse counterpart should allow one to reprove many of the results mentioned in this section in a unified way.

5.3. Combinatorial theorems relative to a pseudorandom set. While this survey has focused on combinatorial theorems relative to random sets, analogous questions may also be asked for pseudorandom sets. Much of the work in this direction has focused on the combinatorial properties of the class of (p,β)-jumbled graphs. These graphs, introduced by Thomason [121, 122], have the property that if X and Y are vertex subsets, then

$$|e(X,Y) - p|X||Y|| \leq \beta\sqrt{|X||Y|}.$$

As one would expect of a pseudorandom property, the random graph $G_{n,p}$ is itself (p,β)-jumbled. In this case, with high probability, we may take β to be $O(\sqrt{pn})$. This is essentially optimal, that is, there are no (p,β)-jumbled graphs with $\beta = o(\sqrt{pn})$. An explicit example of a jumbled graph is the Paley graph. This is the graph with vertex set $\mathbb{Z}_p$, where p is a prime of the form $4k+1$, and edge set given by joining x and y if and only if their difference is a quadratic residue. This graph is again optimally jumbled with $p = \frac{1}{2}$ and $\beta = O(\sqrt{n})$. For many more examples, we refer the reader to the survey [82].

For (p,β)-jumbled graphs, one is usually interested in questions of the following form: given a graph property $\mathcal{P}$, an integer n and a density p, for what values of β is it the case that a (p,β)-jumbled graph on n vertices satisfies $\mathcal{P}$? To give an example, for any integer $t \geq 3$, there exists $c > 0$ such that if $\beta \leq cp^{t-1}n$ then any (p,β)-jumbled graph on n vertices contains a copy of K_t. For $t = 3$, this condition is known to be tight, as shown by an example of Alon [2].

Very recently, a general method for transferring combinatorial theorems to pseudorandom graphs was found by Fox, Zhao and the author [20]. Though we will not attempt an exhaustive survey, the following sample result is representative.

Theorem 5.6. *For any integer t and any $\epsilon > 0$, there exist positive constants δ and c such that if $\beta \leq cp^t n$ then any (p, β)-jumbled graph G on n vertices has the following property. Any subgraph of G containing at most $\delta p^{\binom{t}{2}} n^t$ copies of K_t may be made K_t-free by deleting at most ϵpn^2 edges.*

That is, we have an extension of the removal lemma to subgraphs of pseudorandom graphs. Although we have only stated this result for cliques, there is also a more general statement that applies to all graphs. Moreover, with similar conditions on β, it is possible to prove analogues of many different combinatorial statements. For example, the (K_t, r)-Ramsey property and (K_t, ϵ)-Turán property both hold in pseudorandom graphs with $\beta \leq cp^t n$.

Unfortunately, there is still a gap in these results, even for triangles. For $t = 3$, Theorem 5.6 (which in this case was first proved by Kohayakawa, Rödl, Schacht and Skokan [77]) says that if $\beta \leq cp^3 n$ then the triangle removal lemma holds for subgraphs of a (p, β)-jumbled graph on n vertices. However, it may well be the case that $\beta \leq cp^2 n$ is sufficient. If true, Alon's example would imply that such a result was optimal.

The method of [20] was extended to hypergraphs in [21], under a different type of pseudorandomness hypothesis (though see also [22]). This result was then used to prove a pseudorandom analogue of Szemerédi's theorem. Such a result was a key ingredient in Green and Tao's proof that the primes contain arbitrarily long arithmetic progressions. Their original result states that if a subset of the integers satisfies two pseudorandomness conditions, the linear forms condition and the correlation condition, then it is (k, δ)-Szemerédi. Our results allow one to remove the correlation condition from this statement. Due to space constraints, we are unable to say more here. However, we refer the reader to [23] for further details.

Acknowledgements. This research was supported by a Royal Society University Research Fellowship. I would like to thank Jacob Fox, Tim Gowers, Rob Morris, Wojtek Samotij, Mathias Schacht, Benny Sudakov and Yufei Zhao for helpful discussions on the topics in this paper. Particular thanks are due to Wojtek Samotij for a number of detailed comments on the manuscript.

References

[1] P. Allen, J. Böttcher, H. Hàn, Y. Kohayakawa, and Y. Person, *Blow-up lemmas for sparse graphs*, preprint.

[2] N. Alon, *Explicit Ramsey graphs and orthonormal labellings*, Electron. J. Combin. **1** (1994), Research paper 12, 8pp.

[3] N. Alon, R. A. Duke, H. Lefmann, V. Rödl, and R. Yuster, *The algorithmic aspects of the regularity lemma*, J. Algorithms **16** (1994), 80–109.

[4] L. Babai, M. Simonovits and J. Spencer, *Extremal subgraphs of random graphs*, J. Graph Theory **14** (1990), 599–622.

[5] J. Balogh, C. Lee, and W. Samotij, *Corrádi and Hajnal's theorem for sparse random graphs*, Combin. Probab. Comput. **21** (2012), 23–55.

[6] J. Balogh, R. Morris, and W. Samotij, *Independent sets in hypergraphs*, preprint.

[7] J. Balogh, R. Morris, W. Samotij, and L. Warnke, *The typical structure of sparse K_{r+1}-free graphs*, preprint.

[8] J. Beck, *On size Ramsey numbers of paths, trees and circuits I*, J. Graph Theory **7** (1983), 115–129.

[9] M. Behrisch, *Random graphs without a short cycle*, Master's thesis, Humboldt Universität zu Berlin, 2002.

[10] B. Bollobás, *Random graphs*, second edition, Cambridge Studies in Advanced Mathematics 73, Cambridge University Press, Cambridge, 2001.

[11] B. Bollobás and A. G. Thomason, *Threshold functions*, Combinatorica **7** (1987), 35–38.

[12] J. Böttcher, Y. Kohayakawa, and A. Taraz, *Almost spanning subgraphs of random graphs after adversarial edge removal*, Combin. Probab. Comput. **22** (2013), 639–683.

[13] J. Böttcher, M. Schacht, and A. Taraz, *Proof of the bandwidth conjecture of Bollobás and Komlós*, Math. Ann. **343** (2009), 175–205.

[14] G. Brightwell, K. Panagiotou, and A. Steger, *Extremal subgraphs of random graphs*, Random Structures Algorithms **41** (2012), 147–178.

[15] S. Burr, *On the computational complexity of Ramsey-type problems*, in Mathematics of Ramsey theory, 46–52, Algorithms Combin., 5, Springer, Berlin, 1990.

[16] V. Chvátal, V. Rödl, E. Szemerédi, and W. T. Trotter Jr, *The Ramsey number of a graph with bounded maximum degree*, J. Combin. Theory Ser. B **34** (1983), 239–243.

[17] D. Conlon, *A new upper bound for diagonal Ramsey numbers*, Ann. of Math. **170** (2009), 941–960.

[18] D. Conlon and J. Fox, *Graph removal lemmas*, in Surveys in Combinatorics 2013, London Math. Soc. Lecture Note Ser., Vol. 409, 1–50, Cambridge University Press, Cambridge, 2013.

[19] D. Conlon, J. Fox, and B. Sudakov, *On two problems in graph Ramsey theory*, Combinatorica **32** (2012), 513–535.

[20] D. Conlon, J. Fox, and Y. Zhao, *Extremal results in sparse pseudorandom graphs*, Adv. Math. **256** (2014), 206–290.

[21] D. Conlon, J. Fox, and Y. Zhao, *A relative Szemerédi theorem*, submitted.

[22] ———, *Linear forms from the Gowers uniformity norm*, unpublished note.

[23] ———, *The Green-Tao theorem: an exposition*, submitted.

[24] D. Conlon and W. T. Gowers, *Combinatorial theorems in sparse random sets*, submitted.

[25] ———, *An upper bound for Folkman numbers*, submitted.

[26] D. Conlon, W. T. Gowers, W. Samotij, and M. Schacht, *On the KŁR conjecture in random graphs*, to appear in Israel J. Math.

[27] B. DeMarco and J. Kahn, *Mantel's theorem in random graphs*, preprint.

[28] ———, *Turán's theorem in random graphs*, preprint.

[29] G. A. Dirac, *Some theorems on abstract graphs*, Proc. London Math. Soc. **2** (1952), 69–81.

[30] P. Erdős, *Some remarks on the theory of graphs*, Bull. Amer. Math. Soc. **53** (1947), 292–294.

[31] P. Erdős, P. Frankl, and V. Rödl, *The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent*, Graphs Combin. **2** (1986), 113–121.

[32] P. Erdős and A. Rényi, *On random graphs I*, Publ. Math. Debrecen **6** (1959), 290–297.

[33] ———, *On the evolution of random graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl. **5** (1960), 17–61.

[34] P. Erdős and M. Simonovits, *A limit theorem in graph theory*, Studia Sci. Math. Hungar. **1** (1966), 51–57.

[35] ———, *Supersaturated graphs and hypergraphs*, Combinatorica **3** (1983), 181–192.

[36] P. Erdős and A. H. Stone, *On the structure of linear graphs*, Bull. Amer. Math. Soc. **52** (1946), 1087–1091.

[37] P. Erdős and G. Szekeres, *A combinatorial problem in geometry*, Compos. Math. **2** (1935), 463–470.

[38] J. Folkman, *Graphs with monochromatic complete subgraphs in every edge coloring*, SIAM J. Appl. Math. **18** (1970), 19–24.

[39] J. Fox, *A new proof of the graph removal lemma*, Ann. of Math. **174** (2011), 561–579.

[40] P. Frankl and V. Rödl, *Large triangle-free subgraphs in graphs without K_4*, Graphs Combin. **2** (1986), 135–144.

[41] E. Friedgut, *Sharp thresholds of graph properties, and the k-sat problem*, J. Amer. Math. Soc. **12** (1999), 1017–1054.

[42] E. Friedgut, H. Hàn, Y. Person, and M. Schacht, *A sharp threshold for van der Waerden's theorem in random subsets of $\mathbb{Z}/n\mathbb{Z}$*, preprint.

[43] E. Friedgut and M. Krivelevich, *Sharp thresholds for Ramsey properties of random graphs*, Random Structures Algorithms **17** (2000), 1–19.

[44] E. Friedgut, V. Rödl, A. Ruciński, and P. Tetali, *A sharp threshold for random graphs with a monochromatic triangle in every edge coloring*, Mem. Amer. Math. Soc. **179** (2006), no. 845, vi + 66 pp.

[45] E. Friedgut, V. Rödl, and M. Schacht, *Ramsey properties of discrete random structures*, Random Structures Algorithms **37** (2010), 407–436.

[46] Z. Füredi, *Random Ramsey graphs for the four-cycle*, Discrete Math. **126** (1994), 407–410.

[47] ———, *Extremal hypergraphs and combinatorial geometry*, in Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), 1343–1352, Birkhäuser, Basel, 1995.

[48] H. Furstenberg, *Ergodic behaviour of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.

[49] S. Gerke, Y. Kohayakawa, V. Rödl, and A. Steger, *Small subsets inherit sparse ϵ-regularity*, J. Combin. Theory Ser. B **97** (2007), 34–56.

[50] S. Gerke, M. Marciniszyn, and A. Steger, *A probabilistic counting lemma for complete graphs*, Random Structures Algorithms, **31** (2007), 517–534.

[51] S. Gerke, H. J. Prömel, T. Schickinger, A. Steger, and A. Taraz, *K_4-free subgraphs of random graphs revisited*, Combinatorica **27** (2007), 329–365.

[52] S. Gerke and A. Steger, *The sparse regularity lemma and its applications*, in Surveys in Combinatorics 2005, London Math. Soc. Lecture Note Ser., Vol. 327, 227–258, Cambridge University Press, Cambridge, 2005.

[53] S. Gerke, A. Steger, and T. Schickinger, *K_5-free subgraphs of random graphs*, Random Structures Algorithms **24** (2004), 194–232.

[54] W. T. Gowers, *A new proof of Szemerédi's theorem*, Geom. Funct. Anal. **11** (2001), 465–588.

[55] ———, *Hypergraph regularity and the multidimensional Szemerédi theorem*, Ann. of Math. **166** (2007), 897–946.

[56] ———, *Decompositions, approximate structure, transference, and the Hahn-Banach theorem*, Bull. London Math. Soc. **42** (2010), 573–606.

[57] R. L. Graham, V. Rödl, and A. Ruciński, *On Schur properties of random subsets of integers*, J. Number Theory **61** (1996), 388–408.

[58] ———, *On graphs with linear Ramsey numbers*, J. Graph Theory **35** (2000), 176–192.

[59] R. L. Graham, B. L. Rothschild, and J. H. Spencer, *Ramsey theory*, second edition, John Wiley & Sons, 1990.

[60] B. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions*, Ann. of Math. **167** (2008), 481–547.

[61] L. Gugelmann, *Ramsey properties of random graphs and hypergraphs*, PhD thesis, ETH Zürich, 2013.

[62] L. Gugelmann, Y. Person, A. Steger, and H. Thomas, *A randomized version of Ramsey's theorem*, Random Structures Algorithms **41** (2012), 488–505.

[63] P. E. Haxell, Y. Kohayakawa, and T. Łuczak, *Turán's extremal problem in random graphs: forbidding even cycles*, J. Combin. Theory Ser. B **64** (1995), 273–287.

[64] _____, *Turán's extremal problem in random graphs: forbidding odd cycles*, Combinatorica **16** (1996), 107–122.

[65] H. Huang, C. Lee, and B. Sudakov, *Bandwidth theorem for random graphs*, J. Combin. Theory Ser. B **102** (2012), 14–37.

[66] S. Janson, *Poisson approximation for large deviations*, Random Structures Algorithms **1** (1990), 221–229.

[67] S. Janson, T. Łuczak, and A. Ruciński, *Random graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York, 2000.

[68] P. Keevash, *Hypergraph Turán Problems*, in Surveys in Combinatorics 2011, London Math. Soc. Lecture Note Ser., Vol. 392, 83–140, Cambridge University Press, Cambridge, 2011.

[69] D. J. Kleitman and K. J. Winston, *On the number of graphs without* 4*-cycles*, Discrete Math. **41** (1982), 167–172.

[70] Y. Kohayakawa, *Szemerédi's regularity lemma for sparse graphs*, in Foundations of computational mathematics (Rio de Janeiro, 1997), 216–230, Springer, Berlin, 1997.

[71] Y. Kohayakawa and B. Kreuter, *Threshold functions for asymmetric Ramsey properties involving cycles*, Random Structures Algorithms **11** (1997), 245–276.

[72] Y. Kohayakawa, T. Łuczak, and V. Rödl, *Arithmetic progressions of length three in subsets of a random set*, Acta Arith. **75** (1996), 133–163.

[73] _____, *On* K_4*-free subgraphs of random graphs*, Combinatorica **17** (1997), 173–213.

[74] Y. Kohayakawa and V. Rödl, *Regular pairs in sparse random graphs I*, Random Structures Algorithms **22** (2003), 359–434.

[75] _____, *Szemerédi's regularity lemma and quasi-randomness*, in Recent advances in algorithms and combinatorics, CMS Books Math./Ouvrages Math. SMC, 11, 289–351, Springer, New York, 2003.

[76] Y. Kohayakawa, V. Rödl, and M. Schacht, *The Turán theorem for random graphs*, Combin. Probab. Comput. **13** (2004), 61–91.

[77] Y. Kohayakawa, V. Rödl, M. Schacht, and J. Skokan, *On the triangle removal lemma for subgraphs of sparse pseudorandom graphs*, in An irregular mind (Szemerédi is 70), Bolyai Soc. Math. Stud., Vol. 21, 359–404, Springer, Berlin, 2010.

[78] Y. Kohayakawa, V. Rödl, M. Schacht, and E. Szemerédi, *Sparse partition universal graphs for graphs of bounded degree*, Adv. Math. **226** (2011), 5041–5065.

[79] Y. Kohayakawa, M. Schacht, and R. Spöhel, *Upper bounds on probability thresholds for asymmetric Ramsey properties*, Random Structures Algorithms **44** (2014), 1–28.

[80] J. Komlós, G. Sárközy, and E. Szemerédi, *Blow-up lemma*, Combinatorica **17** (1997), 109–123.

[81] J. Komlós, A. Shokoufandeh, M. Simonovits, and E. Szemerédi, *The regularity lemma and its applications in graph theory*, in Theoretical aspects of computer science (Tehran, 2000), Lecture Notes in Comput. Sci., Vol. 2292, 84–112, Springer, Berlin, 2002.

[82] M. Krivelevich and B. Sudakov, *Pseudo-random graphs*, in More sets, graphs and numbers, Bolyai Soc. Math. Stud., Vol. 15, 199–262, Springer, Berlin, 2006.

[83] D. Kühn and D. Osthus, *Embedding large subgraphs into dense graphs*, in Surveys in Combinatorics 2009, London Math. Soc. Lecture Note Ser., Vol. 365, 137–167, Cambridge University Press, Cambridge, 2009.

[84] C. Lee and B. Sudakov, *Dirac's theorem for random graphs*, Random Structures Algorithms **41** (2012), 293–305.

[85] T. Łuczak, *Randomness and regularity*, in International Congress of Mathematicians, Vol. III, 899–909, Eur. Math. Soc., Zürich, 2006.

[86] T. Łuczak, A. Ruciński, and B.Voigt, *Ramsey properties of random graphs*, J. Combin. Theory Ser. B **56** (1992), 55–68.

[87] M. Marciniszyn, J. Skokan, R. Spöhel, and A. Steger, *Asymmetric Ramsey properties of random graphs involving cliques*, Random Structures Algorithms **34** (2009), 419–453.

[88] B. Nagle, V. Rödl, and M. Schacht, *The counting lemma for regular k-uniform hypergraphs*, Random Structures Algorithms **28** (2006), 113–179.

[89] R. Nenadov and A. Steger, *A short proof of the random Ramsey theorem*, to appear in Combin. Probab. Comput.

[90] J. Nešetřil and V. Rödl, *The Ramsey property for graphs with forbidden complete subgraphs*, J. Combin. Theory Ser. B **20** (1976), 243–249.

[91] _____, *Simple proof of the existence of restricted Ramsey graphs by means of a partite construction*, Combinatorica **1** (1981), 199–202.

[92] R. Rado, *Note on combinatorial analysis*, Proc. London Math. Soc. **48** (1941), 122–160.

[93] F. P. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. **30** (1930), 264–286.

[94] Omer Reingold, Luca Trevisan, Madhur Tulsiani, and Salil Vadhan, *Dense Subsets of Pseudorandom Sets*, in Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science, 76–85, IEEE Computer Society, Washington DC, 2008.

[95] V. Rödl, *Quasi-randomness and the regularity method in hypergraphs*, preprint.

[96] V. Rödl and A. Ruciński, *Lower bounds on probability thresholds for Ramsey properties*, in Combinatorics, Paul Erdős is Eighty, Vol. 1, 317–346, Bolyai Soc. Math. Studies, János Bolyai Math. Soc., Budapest, 1993.

[97] ———, *Random graphs with monochromatic triangles in every edge coloring*, Random Structures Algorithms **5** (1994), 253–270.

[98] ———, *Threshold functions for Ramsey properties*, J. Amer. Math. Soc. **8** (1995), 917–942.

[99] ———, *Rado partition theorem for random subsets of integers*, Proc. London Math. Soc. **74** (1997), 481–502.

[100] ———, *Ramsey properties of random hypergraphs*, J. Combin. Theory Ser. A **81** (1998), 1–33.

[101] V. Rödl, A. Ruciński, and M. Schacht, *Ramsey properties of random k-partite, k-uniform hypergraphs*, SIAM J. Discrete Math. **21** (2007), 442–460.

[102] ———, *Ramsey properties of random graphs and Folkman numbers*, preprint.

[103] V. Rödl and M. Schacht, *Regularity lemmas for graphs*, in Fete of combinatorics and computer science, Bolyai Soc. Math. Stud., Vol. 20, 287–325, János Bolyai Math. Soc., Budapest, 2010.

[104] V. Rödl and J. Skokan, *Regularity lemma for uniform hypergraphs*, Random Structures Algorithms **25** (2004), 1–42.

[105] K. F. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 104–109.

[106] I. Z. Ruzsa and E. Szemerédi, *Triple systems with no six points carrying three triangles*, in Combinatorics (Keszthely, 1976), Vol. II, 939–945, Colloq. Math. Soc. János Bolyai, 18, North-Holland, Amsterdam-New York, 1978.

[107] W. Samotij, *Stability results for random discrete structures*, Random Structures Algorithms **44** (2014), 269–289.

[108] A. A. Sapozhenko, *On the number of connected subsets with given cardinality of the boundary in bipartite graphs (in Russian)*, Metody Diskret. Analiz. **45** (1987), 42–70.

[109] ———, *On the number of independent sets in extenders*, Discrete Math. Appl. **11** (2001), 155–161.

[110] ———, *Independent sets in quasi-regular graphs*, European J. Combin. **27** (2006), 1206–1210.

[111] D. Saxton and A. Thomason, *Hypergraph containers*, preprint.

[112] M. Schacht, *Extremal results for discrete random structures*, preprint.

[113] A. Scott, *Szemerédi's regularity lemma for matrices and sparse graphs*, Combin. Probab. Comput. **20** (2011), 455–466.

[114] M. Simonovits, *A method for solving extremal problems in graph theory, stability problems*, in Theory of Graphs (Proc. Colloq., Tihany, 1966), 279–319, Academic Press, New York, 1968.

[115] J. Spencer, *Ramsey's theorem – a new lower bound*, J. Combin. Theory Ser. A **18** (1975), 108–115.

[116] B. Sudakov and V. H. Vu, *Local resilience of graphs*, Random Structures Algorithms **33** (2008), 409–433.

[117] T. Szabó and V. H. Vu, *Turán's theorem in sparse random graphs*, Random Structures Algorithms **23** (2003), 225–234.

[118] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 199–245.

[119] ———, *Regular partitions of graphs*, in Problèmes combinatoires et théorie des graphes (Orsay 1976), Colloq. Internat. CNRS, 260, 399–401, CNRS, Paris, 1978.

[120] T. Tao, *A variant of the hypergraph removal lemma*, J. Combin. Theory Ser. A **113** (2006), 1257–1280.

[121] A. G. Thomason, *Pseudorandom graphs*, in Random graphs '85 (Poznań, 1985), North-Holland Math. Stud., Vol. 144, 307–331, North-Holland, Amsterdam, 1987.

[122] ———, *Random graphs, strongly regular graphs and pseudorandom graphs*, in Surveys in Combinatorics 1987, London Math. Soc. Lecture Note Ser., Vol. 123, 173–195, Cambridge University Press, Cambridge, 1987.

[123] Paul Turán, *Egy gráfelméleti szélsőértékfeladatról*, Mat. Fiz. Lapok **48** (1941), 436–452.

[124] B. L. van der Waerden, *Beweis einer Baudetschen Vermutung*, Nieuw. Arch. Wisk. **15** (1927), 212–216.

Mathematical Institute, Oxford OX2 6GG, United Kingdom.
E-mail: david.conlon@maths.ox.ac.uk

The graph regularity method: variants, applications, and alternative methods

Jacob Fox

Abstract. Szemerédi's regularity lemma is one of the most powerful tools in graph theory, with many applications in combinatorics, number theory, discrete geometry, and theoretical computer science. Roughly speaking, it says that every large graph can be partitioned into a small number of parts such that the bipartite subgraph between almost all pairs of parts is random-like. Several variants of the regularity lemma have since been established with many further applications. This survey discusses recent progress in understanding the quantitative aspects of these lemmas and their applications, as well as recent progress in developing a sparse regularity method.

Mathematics Subject Classification (2010). 05C35, 05C65, 05D10, 05D40.

Keywords. Regularity lemma, Ramsey theory, extremal combinatorics, probabilistic methods.

1. Introduction

Much of the world can be described as graphs, consisting of discrete elements (called vertices) with connections between certain pairs of them (called edges). Some examples described in the recent book by Lovász [91] include

- The Internet with computers connected by links;
- The World Wide Web with webpages and hyperlinks;
- Social networks like Facebook with users and friendships;
- Chemical networks like imperfect crystals with atoms and chemical bonds;
- Biological networks like the brain with neurons and synapses;
- Engineered networks like integrated circuits with transistors and wires.

Understanding the structure of these graphs can yield critical insights on topics ranging from the spread of diseases to the properties of complex crystals. However, it is often difficult to analyze extremely large networks; each of the examples given above has over a billion vertices. This major practical challenge is related to exciting developments in combinatorics and theoretical computer science.

While graph theory is an old subject with a history that goes as far back as Euler, an important modern direction of research is developing mathematical tools for studying very large graphs. Central to this area is a powerful result of Szemerédi [123] known as the

Proceedings of the International Congress of Mathematicians, Seoul, 2014

regularity lemma. The regularity lemma provides a rough structural description of all large graphs. It roughly states that the vertices of any graph can be partitioned into a bounded number of parts such that the edges between almost every pair of parts behaves in a random-like fashion. This result created a paradigm shift in how we view and study graphs, and it has become a central tool in discrete mathematics with diverse applications in mathematics and computer science.

Szemerédi [122] used an early variant of the regularity lemma in his proof of the celebrated Erdős-Turán conjecture (now known as Szemerédi's theorem) that every subset of the integers of positive upper density contains arbitrarily long arithmetic progressions. The regularity lemma (see the surveys [83], [105], [106]) has since become a central tool in extremal combinatorics, with many applications in number theory, graph theory, theoretical computer science, and discrete geometry.

The regularity lemma requires a few definitions to properly state it. For vertex subsets X, Y of a graph G, $e(X,Y)$ denotes the number of pairs in $X \times Y$ that are edges, and the density $d(X,Y) = \frac{e(X,Y)}{|X||Y|}$ is the proportion of pairs in $X \times Y$ that are edges. A pair of vertex subsets X, Y is *ϵ-regular* if for all $X' \subset X$ and $Y' \subset Y$ with $|X'| \geq \epsilon|X|$ and $|Y'| \geq \epsilon|Y|$, we have $|d(X',Y') - d(X,Y)| < \epsilon$. Thus, the edges between an ϵ-regular pair with ϵ small are uniformly distributed across large subsets. Let $P : V = V_1 \cup \ldots \cup V_k$ be a vertex partition of a graph $G = (V,E)$. The partition P is *equitable* if $|V_i| = \lfloor |V|/k \rfloor$ or $\lceil |V|/k \rceil$ for $1 \leq i \leq k$. The partition P is *ϵ-regular* if all but at most ϵk^2 pairs of parts are ϵ-regular.

Theorem 1.1 (Szemerédi's regularity lemma). *For each $\epsilon > 0$ there is $K(\epsilon)$ such that every graph has an equitable ϵ-regular partition into at most $K(\epsilon)$ parts.*

One major drawback of applying the regularity lemma is that the number of parts $K(\epsilon)$ is an exponential tower of twos of height polynomial in $1/\epsilon$. Unfortunately, this yields weak and impractical quantitative estimates in its various applications. For almost two decades, there was hope that the bound could be substantially improved; however, in 1997, Gowers [61] used a probabilistic construction to show that a tower-type bound is indeed necessary. This work was called a 'tour de force' in the Fields Medal citation for Gowers. Very recently, L. M. Lovász and the author [53] gave a tight lower bound construction for the order on the tower height in a version of the regularity lemma. The proof reverse engineers Szemerédi's upper bound argument and shows that it is essentially best possible. The proof of this regularity lemma and the lower bound construction are discussed in the next section.

Due to the weak quantitative bounds given by the regularity lemma, it is quite desirable to find alternative methods which give better quantitative bounds. We describe one such method, a powerful probabilistic technique known as dependent random choice, and some of its many applications in Section 3. In Section 4, we discuss two additional methods, higher order Fourier analysis and the greedy embedding method. Also in this section we discuss the quantitative aspects of the graph removal lemma, one of the most important applications of the regularity lemma, and some of its extensions. In Section 5, we describe a number of variants of the regularity lemma and their applications. In particular, we describe the sparse regularity method, which extends many classical results to sparse graphs, and recent work by Conlon, Zhao, and the author [27] giving a strengthening of the relative Szemerédi theorem, which simplifies the proof of the Green-Tao theorem [69] that the primes contain arbitrarily long arithmetic progressions; see the recent exposition of the proof [28]. Unfortunately, due to space limitations, many exciting developments related to the regularity method such as

the blow-up lemma and the absorption method are left uncovered.

2. Regularity lemma bounds

We next describe more precisely a version of Szemerédi's regularity lemma. This version was first formulated by Lovász and Szegedy [92], and can easily be shown to be equivalent to the original version of Szemerédi [123]. For a pair of vertex subsets X and Y of a graph G, let $e(X,Y)$ be the number of ordered pairs of vertices $(x,y) \in X \times Y$ that have an edge between them in the graph. Let $d(X,Y) = \frac{e(X,Y)}{|X||Y|}$ be the edge density between X and Y. The *irregularity* of the pair X, Y is defined to be

$$\operatorname{irreg}(X,Y) = \max_{U \subset X, W \subset Y} \big|e(U,W) - |U||W|d(X,Y)\big|.$$

This is a value between 0 and $|X||Y|$. If this is a small fraction of $|X||Y|$, then the edge distribution between X and Y is quite uniform, or random-like. The *irregularity* of a partition $\mathcal{P}$ of the vertex set of G is defined to be

$$\operatorname{irreg}(\mathcal{P}) = \sum_{X,Y \in \mathcal{P}} \operatorname{irreg}(X,Y).$$

Szemerédi's regularity lemma, as stated in [92], is as follows.

Theorem 2.1. *For any $\epsilon > 0$, there is a (least) $M(\epsilon)$ such that any graph $G = (V,E)$ has a vertex partition into at most $M(\epsilon)$ parts with irregularity at most $\epsilon|V|^2$.*

It is easy to show that the two versions of the regularity lemma in Theorem 1.1 and 2.1 are equivalent up to a polynomial change in ϵ. We present here the standard proof of the regularity lemma (as in Theorem 2.1). It shows that $M(\epsilon)$ is at most an exponential tower of twos of height $O(\epsilon^{-2})$.

The key idea that makes the proof work is to use a density increment argument with the *mean square density*. Let $G = (V,E)$ be a graph and let $\mathcal{P}$ be a vertex partition into parts $V_1, V_2, \ldots, V_k$. The mean square density of the partition $\mathcal{P}$ is defined to be

$$q(\mathcal{P}) := \sum_{1 \le i,j \le k} q(V_i, V_j),$$

where $q(V_i,V_j) = \frac{|V_i||V_j|}{|V|^2} d(V_i,V_j)^2$. Since the mean square density is a weighted average of numbers between 0 and 1, every partition $\mathcal{P}$ satisfies $0 \le q(\mathcal{P}) \le 1$. It follows from the Cauchy-Schwarz inequality that if $\mathcal{P}'$ is a refinement of $\mathcal{P}$, then $q(\mathcal{P}') \ge q(\mathcal{P})$. An important observation in the proof is that one can obtain a better inequality if $\mathcal{P}$ is not ϵ-regular and $\mathcal{P}'$ is obtained in a certain way.

For partitions $\mathcal{Q}_X$ and $\mathcal{Q}_Y$ of X and Y, let $q(\mathcal{Q}_X, \mathcal{Q}_Y) = \sum_{U \in \mathcal{Q}_X, W \in \mathcal{Q}_Y} q(U,W)$.

Lemma 2.2. *Suppose $\mathcal{P}$ is a partition with k parts which is not ϵ-regular. Then there is a refinement $\mathcal{Q}$ of $\mathcal{P}$ with at most $k2^{k+1}$ parts and $q(\mathcal{Q}) \ge q(\mathcal{P}) + \epsilon^2$.*

Proof. Let $X, Y \in \mathcal{P}$ and consider the partitions $P_{XY} : X = X_Y^0 \cup X_Y^1$ and $P_{YX} : Y = Y_X^0 \cup Y_X^1$ with

$$\operatorname{irreg}(X,Y) = \left|e(X_Y^0, Y_X^0) - d(X,Y)|X_Y^0||Y_X^0|\right|.$$

Let $\mathcal{Q}_X$ be the partition of X which is the common refinement of all P_{XY} with $Y \in \mathcal{P}$, and $\mathcal{Q}$ be the union of the $\mathcal{Q}_X$ with $X \in \mathcal{P}$, so $\mathcal{Q}$ is a vertex partititon which refines $\mathcal{P}$. As P_{XY} has two parts for each $Y \neq X$, and P_{XX} gives a partition of X into four parts, we obtain $|\mathcal{Q}_X| \leq 2^{k-1}4 = 2^{k+1}$ and $|\mathcal{Q}| \leq k2^{k+1}$. To complete the proof, we have

$$\begin{aligned}
q(\mathcal{Q}) - q(\mathcal{P}) &= \sum_{X,Y} q(\mathcal{Q}_X, \mathcal{Q}_Y) - q(X,Y) \\
&\geq \sum_{X,Y} q(P_{XY}, P_{YX}) - q(X,Y) \\
&= \sum_{X,Y} \sum_{U \in P_{XY}, W \in P_{YX}} \frac{|U||W|}{|V|^2} \left(d(U,W) - d(X,Y)\right)^2 \\
&\geq \sum_{X,Y} \frac{|X_Y^0||Y_X^0|}{|V|^2} \left(d(X_Y^0, Y_X^0) - d(X,Y)\right)^2 \\
&= \sum_{X,Y} \frac{1}{|X_Y^0||Y_X^0||V|^2} \operatorname{irreg}(X,Y)^2 \\
&\geq \sum_{X,Y} \frac{1}{|X||Y||V|^2} \operatorname{irreg}(X,Y)^2 \\
&\geq \left(\sum_{X,Y} \frac{\operatorname{irreg}(X,Y)}{|V|^2} \right)^2,
\end{aligned}$$

where the inequalities are either trivial or by the Cauchy-Schwarz inequality. □

Proof of Szemerédi's regularity lemma. By repeatedly applying the above lemma, starting with the trivial partition $\mathcal{P}_0 = \{V\}$ with one part, we obtain a sequence of refinements $\mathcal{P}_0, \mathcal{P}_1, \ldots$, such that, letting k_i denote the number of parts of $\mathcal{P}_i$, we have $k_0 = 1$, $k_{i+1} \leq k_i 2^{k_i+1}$, and $q(\mathcal{P}_{i+1}) \geq q(\mathcal{P}_i) + \epsilon^2$. As the mean square density must lie in $[0,1]$ and increments by at least ϵ^2 at each step, this process must stop within ϵ^{-2} steps, yielding an ϵ-regular partition $\mathcal{P}_t$ with $t \leq \epsilon^{-2}$ whose number of parts k_t is at most a tower of twos of height $\epsilon^{-2} + O(1)$, completing the proof of the regularity lemma. □

There are a few ways one can improve the constant factor in the above proof. First, note that any partition refines the trivial partition with one part, which has mean square density d^2, where $d = d(V,V)$ is the edge density of the graph, and is refined by the partition into singletons, which has mean square density d. Thus, the interval of possible values for the mean square density is $[d^2, d]$, which has length at most $d - d^2 \leq 1/4$. This improves the number of steps by a factor 4 in the above argument, and hence the tower height by a factor 4. A slightly more careful argument as done in [53] shows that the mean square density actually goes up by at least $4\epsilon^2$ in each iteration, giving a further factor 4 improvement.

Lower bound construction. We next present a lower bound construction which shows that the tower height bound in Szemerédi's regularity lemma as in Theorem 2.1 is tight up to an absolute constant factor. As the proof that this construction indeed gives the lower bound is quite long [53], we will only give some broad idea of how the proof goes.

The key idea is to reverse engineer the upper bound construction. A useful lemma of

Gowers [61] shows that it suffices to construct an edge-weighted graph with weights in $[0,1]$ which requires many parts in any ϵ-regular partition. Here the edge density between two vertex subsets is the sum of the edge weights between the two subsets divided by the product of the orders of the two parts. Gowers' lemma follows from constructing an unweighted graph H from a weighted graph G by taking each pair to be an edge at random with probability equal to its edge weight, independently of the other pairs. A standard application of Chernoff's inequality and the union bound shows that, with high probability, if n is the number of vertices, then the number of edges in H and the number of edges in G between every pair of vertex subsets are within $O(n^{3/2})$ of each other.

It is convenient to construct G as an edge-weighted balanced bipartite graph with parts V_1 and V_2. The edge density between V_1 and V_2 is $1/2$. We construct a sequence of equitable partitions $\mathcal{P}_0, \ldots, \mathcal{P}_s$ of $V = V_1 \cup V_2$ with $s = c\alpha^{-2}$ and $\alpha = C\epsilon$, where c is a small enough positive constant and C is a large enough constant. We start with $\mathcal{P}_0 = \{V_1, V_2\}$. For $0 \le i < s$, the partition $\mathcal{P}_{i+1}$ is a refinement of $\mathcal{P}_i$ for $0 \le i \le s-1$ with exponentially more parts. Note that this is quite similar to the sequence of finer partitions constructed in the proof of Szemerédi's regularity lemma. For each pair of parts $X, Y \in \mathcal{P}_i$ with $X \in V_1$ and $Y \in V_2$, equitably partition $X = X_Y^1 \cup X_Y^2$ and $Y = Y_X^1 \cup Y_X^2$ uniformly at random so that each of these parts are the union of parts from $\mathcal{P}_{i+1}$. We call a pair $X, Y \in \mathcal{P}_i$ with $X \in V_1$ and $Y \in V_2$ *active* if $\alpha \le d(X,Y) \le 1-\alpha$. For each pair $X, Y \in \mathcal{P}_i$ with $X \in V_1$ and $Y \in V_2$, and $a, b \in \{1,2\}$, we have $d(X_Y^a, Y_X^b) = d(X,Y)$ if X, Y is not active, $d(X_Y^a, Y_X^b) = d(X,Y) + \alpha$ if X, Y is active and $a = b$, and $d(X_Y^a, Y_X^b) = d(X,Y) - \alpha$ if X, Y is active and $a \ne b$.

The main claim of the proof, from which the desired lower bound on the number of parts easily follows, is that any ϵ-regular partition $\mathcal{P}$ is close to being a refinement of $\mathcal{P}_s$. This means that almost all vertices are in parts of $\mathcal{P}$ which are mostly contained in a part of $\mathcal{P}_s$. The proof does an amortized counting to show that if a part of $\mathcal{P}$ is not mostly contained in a part of $\mathcal{P}_s$, then it makes a substantial contribution to the irregularity of $\mathcal{P}$. The main difficulty is to account for the possibility that a small portion of a part in $\mathcal{P}$ could break off at each step i.

3. Dependent random choice

There are many problems in extremal combinatorics and Ramsey theory concerning embedding a small or sparse graph into a dense graph. The regularity lemma is quite helpful for such applications, but typically gives weak bounds. One alternative idea for obtaining such an embedding is to first find in a dense graph a large vertex subset U which has the useful property that all (or almost all) small subsets of U have many common neighbors. Then one can use this set U and greedily embed the desired subgraph (assuming it is bipartite) one vertex at a time. An elaboration on this idea can be used if the desired subgraph is not bipartite.

This approach is based on a simple yet surprisingly powerful technique known as *dependent random choice*. Early versions of this technique were proved and applied by various researchers, starting with Gowers [62], Kostochka and Rödl [84], and Sudakov [120]. The basic technique, which is an example of the probabilistic method (see [3]), can be roughly described as follows. We pick within a dense graph G a small vertex subset T uniformly at random. Then the rich set U is simply the set of common neighbors of T. Intuitively,

if some subset of G has only few common neighbors, it is unlikely that all the members of the random set T will be chosen among these neighbors. Hence, we do not expect U to contain any such subset. The survey by Sudakov and the author [58] includes many applications and variants of this basic technique. In this section, we highlight a few of these many applications.

3.1. The Ramsey-Turán problem. One of the earliest applications of the regularity method was an influential result of Szemerédi in Ramsey-Turán theory from 1972. The study of Ramsey-Turán numbers was introduced by Sós [119] and was motivated by the classical theorems of Ramsey and Turán and their connections to geometry, analysis, and number theory; see the nice survey by Simonovits and Sós [117].

Szemerédi's [121] result states that for every $\epsilon > 0$ there is $\delta > 0$ such that every graph on n vertices with at least $(\frac{1}{8}+\epsilon)n^2$ edges contains a K_4 or an independent set of size at least δn. In the other direction, Bollobás and Erdős [10] gave an elegant geometric construction, utilizing the isoperimetric inequality for the high dimensional sphere, of a K_4-free graph on n vertices with independence number $o(n)$ and $(\frac{1}{8} - o(1))n^2$ edges. Roughly speaking, the Bollobás-Erdős graph consists of two disjoint copies of a discretized Borsuk graph, which connect nearly antipodal points on a high dimensional sphere. The bipartite graph between the two disjoint copies is dense and connects points between the two spheres which are close to each other.

Starting with Bollobás and Erdős [10] in 1976, various problems have been asked on estimating the minimum possible independence number of a K_4-free graph in the critical window, when the number of edges is about $\frac{n^2}{8}$. These problems were recently solved by Loh, Zhao, and the author [52]. We next summarize these results.

Sudakov [120] in 2003 used dependent random choice to show that any K_4-free graph with independence number $ne^{-\omega(\sqrt{\log n})}$ has only $o(n^2)$ edges. Indeed, using dependent random choice, in any dense graph, one can find a subset U of $\alpha := ne^{-O(\sqrt{\log n})}$ vertices such that every pair of vertices in U has at least α common neighbors. Either U is an independent set, or U contains an edge e. The vertices of e have at least α common neighbors, which either forms an independent set or contains an edge e', in which case the vertices of e and e' form a K_4. In any case, we get an independent set of order α or a K_4. This bound on the independence number is shown in [52] to be close to best possible. Utilizing the Bollobás-Erdős construction, it is shown that if $\alpha = ne^{-o\left((\log n/\log\log n)^{1/2}\right)}$, then there is a graph on n vertices with $(\frac{1}{8} - o(1))n^2$ edges and independence number at most α. This example shows that dependent random choice gives a close to optimal bound for this problem.

It is shown in [52] using a new variant of the dependent random choice technique that there are constants $c, c' > 0$ such that every K_4-free graph on n vertices with $n^2/8$ edges has independence number at least $cn\log\log n/\log n$, and a construction gives an example with independence number at most $c'n(\log\log n)^{3/2}/(\log n)^{1/2}$.

Finally, a new proof of the Ramsey-Turán result of Szemerédi is given in [52] avoiding all use of regularity which gives the correct quantitative dependence. It is shown that for $(\log\log n)^{3/2}/(\log n)^{1/2} \ll \delta < \delta_0$, every K_4-free graph on n vertices with independence number at most δn has at most $(\frac{1}{8} + \frac{3}{2}\delta)n^2$ edges, and a construction is given with $(\frac{1}{8} + (\frac{1}{3} - o(1))\delta)n^2$ edges.

3.2. The Balog-Szemerédi-Gowers lemma. Gowers gave an early application of dependent random choice in his new proof [62] of Szemerédi's theorem on arithmetic progressions

in dense subsets of the integers. One of the important innovations which Gowers introduced in this work is to use dependent random choice to give much better quantitative bounds for a result of Balog and Szemerédi, whose original proof was based on the regularity lemma. The Balog-Szemerédi-Gowers lemma now has many applications and is one of the most important tools in additive combinatorics.

Let A and B be two sets of integers. The *sumset* $A+B=\{a+b : a\in A, b\in B\}$. For a bipartite graph G with parts A and B and edge set E, define the *partial sumset* $A+_G B=\{a+b : (a,b)\in E\}$. The Balog-Szemerédi lemma states that if $|A|=|B|=n$ and G has cn^2 edges and $|A+_G B|\le Cn$, then one can find $A'\subset A$ and $B'\subset B$ with $|A'|,|B'|\ge c'n$ and $|A'+B'|\le C'n$, where c' and C' depend only on c and C. Due to the use of the regularity lemma, the original proof of the Balog-Szemerédi gave a weak bound on the parameters. Gowers' proof gives a much better bound, showing that $1/c'$ and C' can be bounded by a constant degree polynomial in $1/c$ and C. See the survey by Sudakov and the author [58] for the proof and further discussion.

3.3. Sidorenko's conjecture. Up to this point, we have discussed applications of dependent random choice where we improve quantitative estimates on results for which the regularity lemma was originally used. We will now discuss further advances which would not have been possible without the advent of this new method.

A *homomorphism* from a graph H to a graph G is a mapping $f : V(H)\to V(G)$ such that $(f(u),f(v))$ is an edge of G for each edge (u,v) of H. The *homomorphism density* $t_H(G)$ is the fraction of mappings $f : V(H)\to V(G)$ which are homomorphisms.

A fundamental problem in extremal graph theory asks: how small can $t_H(G)$ be for a graph given that the edge density $t_{K_2}(G)$ of G is p? By taking G to be a random graph with edge density p, we obtain the upper bound of p^m for this problem. The beautiful conjectures of Erdős and Simonovits [116] and Sidorenko [114] suggest that this bound is sharp for bipartite graphs. That is, for any bipartite H there is a $\gamma(H)>0$ such that the number of copies of H in any graph G on N vertices with edge density $p>N^{-\gamma(H)}$ is asymptotically at least the same as in the N-vertex random graph with edge density p. More succintly, for every graph H with m edges, $t_H(G)\ge t_{K_2}(G)^m$. Sidorenko observed that this conjecture has the following equivalent analytic form.

Let μ be the Lebesgue measure on $[0,1]$ and let $h(x,y)$ be a bounded, non-negative, symmetric and measurable function on $[0,1]^2$. Let H be a bipartite graph with vertices $u_1,\dots,u_t$ in the first part and vertices $v_1,\dots,v_s$ in the second part. Let E be the set of pairs (i,j) for which (u_i,v_j) is an edge of H, and $m=|E|$. The analytic form of Sidorenko's conjecture states that

$$\int \prod_{(i,j)\in E} h(x_i,y_j)\,d\mu^{t+s} \ge \left(\int h\,d\mu^2\right)^m.$$

The expression on the left hand side of this inequality is quite common. For example, Feynman integrals in quantum field theory, Mayer integrals in statistical mechanics, and multicenter integrals in quantum chemistry are of this form (see Section 6 of [115] and its references). Naturally, Sidorenko's conjecture has connections to a range of topics, such as matrix theory [5, 8], Markov chains [7, 98], graph limits [90], and quasirandomness. Until a few years ago, Sidorenko's conjecture was known to hold in a few very special cases, e.g., for complete bipartite graphs, trees, even cycles (see [114]) and for cubes [75].

The study of quasirandom graphs was introduced by Thomason [126] and Chung, Gra-

ham, and Wilson [16]. They showed that a large number of interesting graph properties satisfied by random graphs are all equivalent. This idea has been quite influential, leading to the study of quasirandomness in other structures such as hypergraphs [14, 64], groups [65], tournaments, permutations, sequences and sparse graphs (see [15] and it references), and progress on problems in different areas (see, e.g., [18, 64, 65]). It is closely related to Szemerédi's regularity lemma and its recent hypergraph generalization and all proofs of Szemerédi's theorem use some notion of quasirandomness. Finally, there is also the fast-growing study of properties of quasirandom graphs, which has recently attracted lots of attention both in combinatorics and theoretical computer science (see, e.g., [88]).

A sequence $(G_n : n = 1, 2, \ldots)$ of graphs is called *quasirandom* with density p (where $0 < p < 1$) if, for every graph H,

$$t_H(G_n) = p^{|E(H)|} + o(1). \tag{3.1}$$

This property is equivalent to many other properties shared by random graphs. One such property is that the edge density in any vertex subset of linear cardinality is $p + o(1)$. A surprising fact, proved in [16], is that it is enough that (3.1) holds for $H = K_2$ and $H = C_4$ for a graph to be quasirandom. That is, a graph with edge density p is quasirandom with density p if the C_4-density is approximately p^4. A question of Chung, Graham, and Wilson [16] which has received considerable attention (see, e.g., [9]) asks for which graphs H is it true that if (3.1) holds for K_2 and H, then the sequence is quasi-random with density p. Such a graph H is called *p-forcing*. The graph H is *forcing* if it is p-forcing for all p. Chung, Graham, and Wilson prove that even cycles C_{2t} and complete bipartite graphs $K_{2,t}$ with $t \geq 2$ are forcing. Skokan and Thoma [118] generalize this result to all complete bipartite graphs $K_{a,b}$ with $a, b \geq 2$.

It is not difficult to show that a forcing graph must be bipartite and have at least one cycle. Skokan and Thoma [118] ask whether these properties characterize the forcing graphs. Conlon, Sudakov, and the author conjecture that the answer is yes and call it the forcing conjecture.

Conjecture 3.1. *A graph H is forcing if and only if it is bipartite and contains a cycle.*

It is not hard to show that the forcing conjecture is stronger than Sidorenko's conjecture.

Using dependent random choice, Conlon, Sudakov, and the author [22] proved Sidorenko's conjecture for a large class of bipartite graphs. The result states that Sidorenko's conjecture holds for any bipartite graph which has a vertex which is complete to the other part. From this result, an approximate version of Sidorenko's conjecture holds for all graphs. Further, they prove the forcing conjecture for any bipartite graph which has two vertices in one part complete to the other part. Extensions of these results to larger families of graphs were obtained recently by Li and Szegedy using an alternative approach [89], and by Kim, Lee, and Lee [77].

3.4. Conjectures of Hajós and Erdős-Fajtlowicz. A *subdivision* of a graph H is any graph formed by replacing edges of H by internally vertex disjoint paths. This is an important notion in graph theory, e.g., the celebrated theorem of Kuratowski uses it to characterize planar graphs. For a graph G, $\sigma(G)$ is the largest integer p such that G contains a subdivision of a complete graph of order p. Further, $\chi(G)$ is the chromatic number of G, and $\omega(G)$ is the clique number of G.

Since the vertices of a clique must receive different colors in a proper coloring, we have $\chi(G) \geq \omega(G)$. A famous conjecture of Hajós from 1961 gives a partial converse to this

fact. It states that $\sigma(G) \geq \chi(G)$. That is, every graph of chromatic number k contains a subdivision of K_k. Dirac [34] proved that this conjecture is true for $k \leq 4$, but in 1979, Catlin [12] disproved the conjecture for $k \geq 7$. By considering a random graph on n vertices, Erdős and Fajtlowicz [40] in 1981 further showed that the conjecture actually fails (and quite strongly) for almost all graphs. They proved that almost all graphs on n vertices satisfy $\chi(G) = \Theta(n/\log n)$ while $\sigma(G) = \Theta(n^{1/2})$. Let $H(n)$ denote the maximum of $\chi(G)/\sigma(G)$ over all n-vertex graphs G. While the Hajós conjecture is equivalent to $H(n) = 1$ for all n, the Erdős-Fajtlowicz result shows that $H(n) \geq cn^{1/2}/\log n$ for some absolute constant $c > 0$. Erdős and Fajtlowicz further conjectured that the random graph is essentially the strongest possible counterexample to the Hajós conjecture in that there is an absolute constant C such that $H(n) \leq Cn^{1/2}/\log n$ for all n. Erdős [39] featured this conjecture in his paper *'On the combinatorial problems I would most like to see solved.'*

The Erdős-Fajtlowicz conjecture was recently proved by Lee, Sudakov, and the author [50]. The proof uses dependent random choice together with several additional tools from extremal graph theory. Dependent random choice is used in the argument to find, in a dense graph G, a large subset U of vertices such that every pair of vertices in U have many paths of length four where the three internal vertices are in $V(G) \setminus U$. The goal is to use this nice subset to help construct a large clique subdivision, say of size s. We find in U a subset S of order s with as many edges as possible. The vertices of S are the vertices of the clique subdivision, the edges in S are used in the clique subdivision, and, using the property of U, we greedily connect by paths of length four each nonadjacent pair of vertices in S to obtain the desired clique subdivision.

3.5. Two extensions of Ramsey's theorem. The Ramsey number $r(k)$ is the minimum n such that every two-coloring of the edges of the complete graph K_n contains a monochromatic K_k. Ramsey's theorem [101] states that $r(k)$ exists for all k. Classical results of Erdős and Szekeres [45] and Erdős [37] demonstrate that $2^{k/2} \leq r(k) \leq 2^{2k}$ for $k \geq 2$. Over the last seven decades, there have been many attempts and several improvements on these bounds (see, e.g., [18]). However, the constant factors in the above exponents have remained unchanged.

The field has naturally stretched in different directions. One such direction is to try to strengthen Ramsey's theorem and guarantee the existence of a monochromatic clique that has some additional structure.

Erdős was interested in the distribution of monochromatic cliques in edge-colorings, and considered the following variant of Ramsey's theorem. For a finite set S of integers greater than one, define its weight $w(S) := \sum_{s \in S} \frac{1}{\log s}$. For a two-edge-coloring of the complete graph on $[2, n] = \{2, \ldots, n\}$, let $f(c)$ be the maximum weight $w(S)$ over all sets $S \subset [2, n]$ which form a monochromatic clique in coloring c. For each integer $n \geq 2$, let $f(n)$ be the minimum of $f(c)$ over all two-edge-colorings c of the complete graph on $[2, n]$. Note that simply applying $r(k) \leq 2^{2k}$ yields only $f(n) \geq \frac{\log n}{2} \frac{1}{\log n} = \frac{1}{2}$.

In his 1981 paper *'On the combinatorial problems I would most like to see solved',* Erdős [39] conjectured that $f(n)$ tends to infinity, and further asked for an accurate estimate of $f(n)$. Rödl [104] verified this conjecture, showing that $f(n) = \Omega(\frac{\log\log\log\log n}{\log\log\log\log\log n})$. In the other direction, a uniform random coloring shows that $f(n) = O(\log\log n)$. Rödl [104] further improved this to $f(n) = O(\log\log\log n)$. Nevertheless, an exponential gap between the lower and upper bound remained. Recently, Conlon, Sudakov, and the author [25] proved that $f(n) = \Theta(\log\log\log n)$. Dependent random choice is an essential ingredient in the

proof.

We next describe Rödl's coloring which shows that $f(n) = O(\log\log\log n)$. Partition $[2,n]$ into $t \approx \log\log n$ intervals, where the ith interval is $[2^{2^{i-1}}, 2^{2^i})$. Two-color the edges within the ith interval so that it contains no monochromatic clique of order 2^{i+1}. This can be done using the fact that the Ramsey number satisfies $r(k) \geq 2^{k/2}$. Thus, the maximum weight of a monochromatic clique within the ith interval is at most $2^{i+1}\frac{1}{\log 2^{2^{i-1}}} = 4$. There is a two-coloring of the edges of K_t with no monochromatic clique of order $2\log t$. Color the edges of the complete bipartite graph between interval i and interval j by the color of the edge (i,j) in this coloring. We obtain a two-edge-coloring of the complete graph on $[2,n]$ such that any monochromatic clique in this coloring has nonempty intersection with less than $2\log t$ intervals. Since each interval can contribute weight at most 4 to such a monochromatic clique, the weight of any monochromatic clique in this coloring is at most $4 \cdot 2\log t = O(\log\log\log n)$.

In [25], we also used dependent random choice to give an exponential improvement on another well-studied problem in this area. Motivated by a problem in model theory, Väänänen asked whether, for any positive integers k and q and any permutation π of $[k-1] = \{1,\dots,k-1\}$, there is a positive integer R such that any q-coloring of the edges of the complete graph on R vertices contains a monochromatic K_k with vertices $a_1 < \ldots < a_k$ such that the consecutive differences $a_2-a_1, a_3-a_2, \dots, a_k-a_{k-1}$ satisfy the same ordering as π. That is, we want a monochromatic clique whose differences between consecutive vertices satisfies a prescribed ordering. The least such positive integer R which works for k, q and all $(k-1)$-permutations π is denoted by $R(k;q)$.

Väänänen's question was popularized by Joel Spencer and answered in the positive by Alon and independently by Erdős, Hajnal, and Pach [42]. Alon's proof (see [97]) uses the Gallai-Witt theorem and gives a weak bound on $R(k;q)$. The proof by Erdős, Hajnal, and Pach uses a compactness argument and gives no bound on $R(k;q)$. Later, Alon, Shelah and Stacey all independently found proofs giving tower-type bounds for $R(k;q)$. A natural conjecture, made by Alon (see [113]), is that $R(k;q)$ should grow exponentially in k. For monotone sequences, this was confirmed by Alon and Spencer. A breakthrough on this problem was obtained by Shelah [113], who proved the double-exponential upper bound $R(k;q) \leq 2^{(q(k+1)^3)^{qk}}$.

Conlon, Sudakov, and the author [25] use dependent random choice to show that $R(k;q) \leq 2^{k^{20q}}$. Thus, for fixed q, $R(k;q)$ grows as a single exponential in a power of k.

3.6. Erdős-Hajnal conjecture. A graph H is an *induced subgraph* of a graph G if there is a one-to-one mapping $f : V(H) \to V(G)$ such that every edge of H maps to an edge of G, and every nonadjacent pair of vertices in H maps to a nonadjacent pair of vertices in G. A graph is *H-free* if it does not contain H as an *induced* subgraph. A basic property of large random graphs is that they almost surely contain any fixed graph H as an induced subgraph. Conversely, there is a general belief that H-free graphs are highly structured. For example, one of the most famous problems in graph theory, the Erdős-Hajnal conjecture [41], is of this sort. It states that every H-free graph on n vertices contains a homogeneous set (i.e., a clique or independent set) of size at least $n^{c(H)}$, where $c(H) > 0$ depends only on H. Erdős and Hajnal proved that such a graph has a homogeneous set of order $2^{c(H)\sqrt{\log n}}$. This is in striking contrast to general graphs on n vertices where one cannot guarantee a homogeneous

set of size larger than logarithmic in n.

The Erdős-Hajnal conjecture has only been solved for a few specific graphs H; see the recent survey by Chudnovsky [13]. An interesting partial result for the general case was obtained by Erdős, Hajnal, and Pach [43]. They show that every H-free graph G with $n \geq 2$ vertices or its complement $\bar{G}$ contains a complete bipartite graph with parts of size $n^{c(H)}$. Sudakov and the author obtained [56] a strengthening of this result which brings it closer to the Erdős-Hajnal conjecture. This result states that any H-free graph on $n \geq 2$ vertices contains a complete bipartite graph with parts of size $n^{c(H)}$ or an independent set of size $n^{c(H)}$.

To get a better understanding of the properties of H-free graphs, it is also natural to ask for an asymmetric version of the Erdős-Hajnal result. In [56] we show that there exists $c(H) > 0$ such that for any H-free graph G on n vertices and n_1, n_2 satisfying $(\log n_1)(\log n_2) \leq c(H) \log n$, G contains a clique of size n_1 or an independent set of size n_2. The proof of both of the above mentioned results from [56] utilize dependent random choice, while the previous techniques are not strong enough to obtain these results.

Using Szemerédi's regularity lemma, Rödl [103] proved the following Ramsey-type result for forbidden induced subgraphs. For each $\epsilon > 0$ and graph H there is $\delta = \delta(\epsilon, H) > 0$ such that every H-free graph on n vertices contains an induced subgraph with at least δn vertices and edge density at most ϵ or at least $1 - \epsilon$. The use of the regularity lemma leads to a weak estimate on δ. Sudakov and the author [55] proved a better bound, showing that we can take $\delta = 2^{-c(H)(\log 1/\epsilon)^2}$. The proof uses the greedy embedding method discussed in Subsection 4.2. The following corollary of this result and a result of Erdős and Szemerédi [46] is also obtained in [55]. Every H-free graph on n vertices contains a homogeneous set of order $c_1 2^{c_2\sqrt{(\log n)/|H|}} \log n$. In addition to implying the Erdős-Hajnal bound, it also implies a result of Prömel and Rödl [100] which shows that either a graph contains an unusually large homogeneous set, or it contains all graphs up to logarithmic size as induced subgraphs. Precisely, it says that for each C there is $c > 0$ such that every graph on n vertices contains as an induced subgraph every graph on at most $c \log n$ vertices or contains a homogeneous set of order $C \log n$.

Sudakov and the author [55] have made the following conjecture which states that we may take $\delta = \epsilon^{c(H)}$ in Rödl's result and would imply the Erdős-Hajnal conjecture.

Conjecture 3.2. *For each $\epsilon > 0$, every H-free graph on n vertices contains an induced subgraph on $\epsilon^{c(H)} n$ vertices with edge density at most ϵ or at least $1 - \epsilon$.*

4. Further alternative methods

In the previous section, we discussed dependent random choice as an alternative to the regularity method. There have also been several quite fruitful alternative methods which have been developed, which we discuss briefly in this section.

4.1. Higher order Fourier analysis. Szemerédi's theorem is one of the most famous applications of the graph regularity method. It states that for each $\delta > 0$ and positive integer k, there is $N(k, \delta)$ such that every subset $A \subset [N]$ with $N \geq N(k, \delta)$ and $|A| \geq \delta N$ contains a k-term arithmetic progression. The original proof using the regularity lemma gives an Ackermann-type bound.

However, the first proof in the case $k = 3$ is due to Roth [108] from 1953 and uses Fourier analysis, giving a much better bound. It is now well-understood that simply using Fourier analysis is not sufficient to count longer arithmetic progressions. Gowers [63] developed higher order Fourier analysis in order to give a new proof of Szemerédi's theorem which gives a much more reasonable bound, namely

$$N(k,\delta) \leq 2^{2^{\delta^{-2^{2^{k+9}}}}}.$$

Higher order Fourier analysis has played a particularly important role in the development of additive combinatorics, and is used by Green and Tao [69] in their proof that the primes contain long arithmetic progressions, and in the work of Green, Tao, and Ziegler [70, 72–74] on linear equations in the primes. For more on higher order Fourier analysis, see the recent book of Tao [124].

4.2. The greedy embedding method. The Ramsey number $r(H)$ of a graph H is the minimum N such that every two-coloring of the edges of the complete graph K_N contains a monochromatic copy of H. As already mentioned, for the complete graph on n vertices, we have $2^{n/2} \leq r(K_n) \leq 2^{2n}$. One direction initiated by Burr and Erdős is to study Ramsey numbers of sparse graphs.

In particular, Burr and Erdős [11] conjectured that for each Δ there is $c(\Delta)$ such that every graph H with maximum degree Δ on n vertices satisfies $r(H) \leq c(\Delta)n$. This conjecture was verified by Chvátal, Rödl, Szemerédi and Trotter [17] in one of the early applications of Szemerédi's regularity lemma. Hence, the Ramsey number of graphs of fixed maximum degree grow only linearly in the number of vertices. Unfortunately, because it uses the regularity lemma, this proof gives a weak upper bound on $c(\Delta)$, namely a tower of twos whose height is exponential in Δ. It has since been an interesting challenge to determine the growth of $c(\Delta)$. Eaton [36], using a weak regularity lemma (see Subsection 5.1), improved the bound on $c(\Delta)$ to double-exponential.

Shortly after, Graham, Rödl and Ruciński [66] proved, by a beautiful method which avoids any use of the regularity lemma, that there exists a constant c for which $c(\Delta) \leq 2^{c\Delta(\log\Delta)^2}$. For bipartite graphs, they [67] improved this bound by removing one logarithmic factor in the exponent. They also proved that there are bipartite graphs with n vertices and maximum degree Δ for which the Ramsey number is at least $2^{c'\Delta}n$.

Using dependent random choice, Conlon [19], and, independently, Sudakov and the author [56] have shown how to remove the logarithmic factor in the exponent for bipartite graphs. More recently, Conlon, Sudakov, and the author [23] extended the greedy embedding method of Graham, Rödl, and Ruciński [66] and improved the bound on $c(\Delta)$ to $2^{c\Delta\log\Delta}$.

Using the hypergraph regularity method, it is shown in [30, 31, 95] that for each k and Δ, there is $c(\Delta,k)$ such that every k-uniform hypergraph on n vertices has Ramsey number at most $c(\Delta,k)n$. This yields Ackermann-type bounds on $c(\Delta,k)$. Extending the dependent random choice technique to hypergraphs, Conlon, Sudakov, and the author [23] proved that $c(\Delta,3) \leq 2^{2^{c\Delta\log\Delta}}$, and, for $k \geq 4$, $c(\Delta,k)$ is at most a tower of height $k-1$ in Δ, which is essentially best possible.

Burr and Erdős also made the stronger conjecture that graphs of bounded *degeneracy* have linear Ramsey numbers. A graph is *d-degenerate* if every subgraph of it has minimum degree at most d. Burr and Erdős conjectured that for each d there is c_d such that every d-degenerate graph H on n vertices has $r(H) \leq c_d n$. Kostochka and Sudakov [85] proved

an upper bound of the form $r(H) \leq n^{1+o(1)}$. The best known bound, due to Sudakov and the author [57], is of the form $r(H) \leq ne^{c(d)\sqrt{\log n}}$. The proofs of these results develop a variant of dependent random choice.

The techniques developed to estimate Ramsey numbers of sparse graphs have also had a variety of other applications. We next briefly describe one more such application. The *induced Ramsey number* $r_{ind}(H)$ is the smallest natural number N for which there is a graph G on N vertices such that in every two-coloring of the edges of G there is an induced monochromatic copy of H. The existence of these numbers was independently proven by Deuber [33], Erdős, Hajnal and Pósa [44] and Rödl [102]. The original proofs give enormous bounds on $r_{ind}(H)$, but it was conjectured by Erdős [38] that the actual values should be more closer to the ordinary Ramsey numbers. In particular, Erdős conjectured that there is a constant c such that every graph H on n vertices satisfies $r_{ind}(H) \leq 2^{cn}$. If true, the complete graph shows that it would be best possible. Despite progress [24, 55, 56, 80, 93] on bounding induced Ramsey numbers, this conjecture is still open. The best known bound is $r_{ind}(H) \leq 2^{cn \log n}$ proved by Conlon, Sudakov and the author [24] using the same method developed in that paper to obtain the best known bound on Ramsey numbers of bounded degree graphs.

We next sketch the regularity-based proof that every graph H on n vertices of maximum degree Δ satisfies $r(H) \leq c(\Delta)n$. Consider a red-blue edge-coloring of the complete graph K_N, where $N = cn$ with c chosen large enough depending only on Δ. Apply Szemerédi's regularity lemma with $\epsilon = C^{-\Delta}$ and obtain an equitable ϵ-regular partition. As all but a small fraction of the pairs of parts are ϵ-regular, we can obtain parts $V_1, \ldots, V_r$ with r being the Ramsey number $r(\Delta + 1)$ such that each pair of parts is ϵ-regular. Consider the red-blue edge-coloring of K_r where (i, j) is red if $d(V_i, V_j) \geq 1/2$ and blue otherwise. By Ramsey's theorem, there are $\Delta + 1$ of the parts, lets call them $U_1, \ldots, U_{\Delta+1}$, such that each pair (U_i, U_j) with $i \neq j$ is ϵ-regular and has density at least $1/2$ in the same color, say red.

Since H has maximum degree Δ, there is a proper coloring $\chi : V(H) \to [\Delta + 1]$. We can then find a red copy of H where the embedding $f(v)$ of each vertex $v \in V(H)$ is in $U_{\chi(v)}$. Suppose the vertices of H are $\{1, \ldots, n\}$. One can greedily find such a red copy one vertex at a time. After step i, we have already picked out the embedding of the first i vertices, $f(1), \ldots, f(i)$, and we have subsets $V_{j,i}$ for $i < j \leq n$ of potential vertices to embed vertex j given the first i vertices have been embedded. We begin with $V_{j,0} = U_{\chi(j)}$ for $1 \leq j \leq n$. The size of $V_{j,i}$ is at least $4^{-d(j,i)}|U_{\chi(j)}| - i$, where $d(j, i)$ denotes the number of neighbors h of j with $h \leq i$. Using that $d(j, i) \leq \Delta$ and the pairs of parts are ϵ-regular with density at least $1/2$, one can pick $f(i + 1)$ from $V_{i+1,i}$ appropriately so that for $j > i + 1$, letting $V_{j,i+1} = V_{j,i} \setminus f(i + 1)$ with j not a neighbor of $i + 1$, and $V_{j,i+1}$ be the red neighborhood of j in $V_{j,i}$ if j is a neighbor of $i + 1$, we can continue the embedding which completes the proof.

The approach of Graham, Rödl, and Rucinski [66] has some similarities to the approach described above, but gets rid of applying regularity. Again, we have a red-blue edge-coloring of K_N. We let $V_{j,0} = V(K_N)$ for $1 \leq j \leq n$. We try to greedily embed a monochromatic red copy of H. At the end of step i we have already embedded $f(1), \ldots, f(i)$, and have potential sets $V_{j,i}$ of vertices to embed the future vertices $j > i$. We would like to keep the property that $|V_{j,i}| \geq (8\Delta)^{-d_{j,i}}N - i$ at each step i and every $j > i$. Note that $|V_{j,i+1}| \geq |V_{j,i}| - 1$ if j is not a neighbor of $i + 1$ as $f(i + 1)$ might be in $V_{j,i}$. If we can guarantee that for each edge $(i + 1, j)$ of H with $j > i + 1$ we have $|V_{j,i+1}| \geq \frac{1}{8\Delta}|V_{j,i}|$, then we can guarantee that each $V_{j,i}$ will be large enough for the above property to hold. If we cannot

find a vertex for $f(i+1)$ at step $i+1$ to continue the embedding, the only reason is that there is $A \subset V_{i+1,i}$ with $|A| \geq \frac{1}{\Delta}|V_{i+1,i}|$ and $B = V_{j,i}$ with j some neighbor of $i+1$ such that the red edge density between A and B is at most $\frac{1}{8\Delta}$. Thus, the blue edge density between A and B is at least $1 - \frac{1}{8\Delta}$.

It would be helpful if instead of having two large subsets with large blue density between them, we had one large subset U with large blue density inside. We could then greedily embed a blue copy of H one vertex at a time. To obtain this, it would suffice to find 4Δ sets of equal size for which blue is very dense between them, and let U be the union of these sets. However, one can iterate the above argument within each subset to either obtain a red copy of H or a pair of large subsets with blue density at least $1 - \frac{1}{8\Delta}$ between them. This iteration loses roughly a factor $\Delta^{-\Delta}$ at each step in the size of the subsets. At each iteration, we double the number of subsets, so we will be done in roughly $\log 4\Delta$ iterations, leading to the bound $c(\Delta) \leq 2^{c\Delta(\log \Delta)^2}$. A simple way to do this iteration process is done in [55], leading to many further applications.

Note the asymmetry between the two colors in the above approach. We either find a large set for which all pairs of large subsets have at least some constant density between them in red, or there is a large subset which is almost complete in blue. In either case, we can greedily embed a monochromatic copy of H.

In [24], Conlon, Sudakov, and the author develop an approach which is symmetric between the two colors and improves the bound on $c(\Delta)$ by removing one of the two logarithmic factors in the exponent.

4.3. Graph removal lemmas. The graph removal lemma is one of the most influential applications of the regularity lemma; see the survey by Conlon and the author [21]. It says that for each $\epsilon > 0$ and graph H on h vertices there is $\delta = \delta(\epsilon, H) > 0$ such that from every graph on n vertices with at most δn^h copies of H one can delete ϵn^2 edges and remove all copies of H. The graph removal lemma has many applications to extremal problems for graphs and hypergraphs, additive combinatorics, discrete geometry, and theoretical computer science. The only known proof used the regularity lemma, leading to weak bounds for the graph removal lemma and its applications. Hence, finding a proof which yields better bounds by avoiding the regularity lemma was a problem of considerable interest and was reiterated by several authors, including Alon, Erdős, Gowers, and Tao.

The author gave a new proof of the graph removal lemma in [47] which gives a bound on $1/\delta$ which is a tower of twos of height logarithmic in $1/\epsilon$. See also [21] for a shorter proof. For comparison, any proof using Szemerédi's regularity lemma would result in a bound which is a tower of twos of height polynomial in $1/\epsilon$. However, there is still a very large gap between our new upper bound for $1/\delta$ and the best known lower bound, which is still slightly superpolynomial in $1/\epsilon$ and based on using the best known lower bound for Roth's theorem. A dramatic improvement of the upper bound would have major consequences in number theory, extremal combinatorics, and computer science, and closing the gap remains an exciting problem.

In certain potential applications of the regularity lemma, it would be helpful if all pairs of parts are ϵ-regular. This is impossible, as the half graph is a simple example in which each equitable partition into k parts yields $\Omega(k)$ irregular pairs. Gowers [61] posed the problem of determining the number of irregular pairs in the regularity lemma. This problem was solved by Conlon and the author [20], showing that there is an absolute constant $\epsilon > 0$ such that for each k there is a graph which requires $\Omega(k^2/\log^* k)$ irregular pairs (i.e., not ϵ-regular) in

every equitable partition into k parts. Here $\log^*$ is the iterated logarithm function. The proof of the regularity lemma shows that this bound on the number of irregular pairs is tight up to the constant factor.

The *induced graph removal lemma* is a result for which one cannot simply apply Szemerédi's regularity lemma due to the possibility of irregular pairs. This result states that for each $\epsilon > 0$ and graph H on h vertices, there is $\delta' = \delta'(\epsilon, H)$ such that every graph on n vertices with at most $\delta' n^h$ induced copies of H can be made induced H-free by adding or deleting at most ϵn^2 edges.

To get around the issue of irregular pairs, Alon, Fischer, Krivelevich, and Szegedy [1] developed the *strong regularity lemma*, see Subsection 5.2. This gives a wowzer-type bound, which is one level higher in the Ackermann hierarchy then the tower function, on $1/\delta$ as a function of $1/\epsilon$. As discussed in Subsection 5.2, such a wowzer bound is indeed necessary for the strong regularity lemma. Addressing a question of Alon on improving this bound, Conlon and the author [20] found a new proof of the induced graph removal lemma which avoids using the strong regularity lemma and gives a tower-type bound. The induced graph removal lemma was extended to the infinite induced removal lemma by Alon and Shapira [4], giving a very general result in graph property testing that natural properties are testable. The wowzer-type bounds were improved to tower-type in these results in graph property testing by Conlon and the author, see [21].

5. Variants of the regularity lemma

In this section, we describe several variants of the regularity lemma and their applications.

5.1. Weak regularity lemmas. In an effort to obtain better bounds in applications, several weak regularity lemmas have been established which are sufficient for certain applications of the regularity method. These typically have single-exponential-type bounds as opposed to the tower-type bound in Szemerédi's regularity lemma, but have the drawback that they have fewer applications.

The most well-known of the weak regularity lemmas is the Frieze-Kannan regularity lemma [59]. It states that for each $\epsilon > 0$ there is $k(\epsilon)$ such that every graph G has a vertex partition $V(G) = V_1 \cup \ldots \cup V_k$ into $k \leq k(\epsilon)$ parts such that

$$\left| e(X,Y) - \sum_{1 \leq i,j \leq k} d(V_i, V_j)|X \cap V_i||Y \cap V_j| \right| \leq \epsilon |V|^2.$$

Unlike Szemerédi's regularity lemma, the Frieze-Kannan regularity lemma does not give us control on the edge density between subsets of parts, but only between large vertex subsets of the graph. However, the Frieze-Kannan regularity lemma is still sufficient for some applications, and the proof gives a bound of $k(\epsilon) = 2^{O(\epsilon^{-2})}$. It follows a similar iterative procedure as done in the proof of Szemerédi's regularity lemma discussed in Section 2, with the mean square density increasing by $\Omega(\epsilon^{-2})$ at each step. The gain comes from the fact that the number of parts is at most a factor four in each step, as opposed to exponentiating the number of parts in each step as in the proof of Szemerédi's regularity lemma. Answering a question of Lovász and Szegedy [92], Conlon and the author [20] gave a construction showing that $k(\epsilon) = 2^{\Omega(\epsilon^{-2})}$. The proof reverse engineers the upper bound proof and has some

similarities to the lower bound construction for the regularity lemma discussed in Section 2.

Another regularity lemma, known as the cylinder regularity lemma, was developed by Duke, Lefmann, Rödl [35]. A *k-cylinder* is a product of k vertex subsets. They show that for any k-partite graph with parts $V_1, \ldots, V_k$, one can partition the complete cylinder $V_1 \times \ldots \times V_k$ into at most $2^{O(\epsilon^{-5}k^2)}$ subcylinders so that all but an ϵ-fraction of the k-tuples are in subcylinders $U_1 \times \cdots \times U_k$ for which each pair (U_i, U_j) is ϵ-regular. They use this lemma to show that for each graph H on k vertices, one can determine the number of copies of H in a graph on n vertices to within an additive approximation of ϵn^k in running time $2^{(k/\epsilon)^{O(1)}} n^{O(1)}$. The running time was recently improved by Grinshpun, Lovász, Zhao, and the author [48] to $\epsilon^{-O(k^2)} n^2$ using an algorithmic cut norm decomposition version of the Frieze-Kannan weak regularity lemma [32].

For some applications of the regularity lemma, it suffices to find just a single ϵ-regular pair, or even a pair of subsets in which each pair of large subsets have some lower bound on the density. The bounds on the size of such a pair was determined by Peng, Rödl, and Ruciński [99]. In other applications, it is sufficient to find a k-tuple of subsets which are pairwise regular, such a k-tuple can be found using the Duke-Lefmann-Rödl cylinder regularity lemma discussed above.

The earliest weak regularity lemma was established by Szemerédi and was used in his proof of Szemerédi's theorem [122]. It states that for each ϵ there is $\ell(\epsilon)$ such that for every bipartite graph with parts U and V, there is a partition $U = U_1 \ldots \cup \ldots U_k$, such that for each U_i, there is a partition of $V = V_{i1} \cup \ldots \cup V_{i\ell_i}$ such that every pair (U_i, V_{ij}) with $1 \leq i \leq k$ and $1 \leq j \leq \ell_i$ is ϵ-regular, where $k, \ell_i \leq \ell(\epsilon)$.

5.2. Strong regularity lemma. Alon, Fischer, Krivelevich, and Szegedy [1] developed the *strong regularity lemma* in order to prove some results that do not follow from directly applying Szemerédi's regularity lemma. The strong regularity lemma is proved by repeated application of Szemerédi's regularity lemma and yields wowzer-type bounds on the number of parts.

The strong regularity lemma states that for each $\epsilon > 0$ and function $f : \mathbb{N} \to (0, 1)$, there is $M = M(\epsilon, f)$ such that every graph G has an equitable partition P and an equitable refinement Q of P with at most M parts such that Q is $f(|P|)$-regular and mean square densities of P and Q are within ϵ of each other, i.e., $q(Q) \leq q(P) + \epsilon$. This last condition essentially says that the edge density between almost all pairs of parts of Q is close to the edge density between the pair of parts of P that they lie in. In [1] they use the strong regularity lemma to prove the induced graph removal lemma. Later, the strong regularity lemma was instrumental in proving very general results stating that natural graph properties are testable; see Subsection 4.3.

Due to the iterated application of the regularity lemma, the number of parts in the strong regularity lemma is of wowzer-type, which is one level higher in the Ackermann hierarchy than the tower function. Conlon and the author [20] and independently Kalyanasundaram and Shapira [76] proved that a wowzer-type bound is indeed necessary.

5.3. Hypergraph regularity method. The regularity method was extended to hypergraphs by Gowers [64] and independently by Nagle, Rödl, Schacht, Skokan [96, 107]. As a consequence, they proved the hypergraph removal lemma. It states that for each k-uniform hypergraph H on h vertices, every k-uniform hypergraph on n vertices with $o(n^h)$ copies of H can be made H-free by removing $o(n^k)$ edges. Szemerédi's theorem and its mult-

dimensional generalization by Furstenberg and Katznelson quickly follow from this result. Many applications of the graph regularity lemma have now been extended to k-uniform hypergraphs by applying the hypergraph regularity method, see [105].

5.4. Sparse regularity method and the primes. Some of the most exciting theoretical and practical problems involve sparse graphs. However, one of the limitations of Szemerédi's regularity lemma is that it is only meaningful for dense graphs. In the 1990s, Kohayakawa and Rödl (see [60, 78, 111]) proved an analogue of Szemerédi's regularity lemma for sparse graphs as part of a general program toward extending extremal results to sparse graphs.

The recent version of Scott [111] removes the assumption that there are no "dense spots". We next give its statement. In a graph with edge density p, a pair of subsets X, Y is (ϵ)-*regular* if, for all $X' \subset X$ and $Y' \subset Y$ with $|X'| \geq \epsilon|X|$ and $|Y'| \geq \epsilon|Y|$, we have $|d(X', Y') - d(X, Y)| < \epsilon p$. Note that the additional p factor gives a tighter regularity condition than the usual condition for sparse graphs. A partition of the vertex set is (ϵ)-*regular* if all but an ϵ-fraction of the pairs of parts are (ϵ)-regular. The sparse regularity lemma then says that for each $\epsilon > 0$ there is $M = M(\epsilon)$ such that every graph has an (ϵ)-regular partition into at most M parts. The proof is similar to the usual proof of the regularity lemma. The key difference is to use a different density function to increment. Instead of using the mean square density, it uses the mean f-density, where f is a convex function satisfying $f(x) = x^2$ for $x \leq C$, and $f(x)$ is linear for $x > C$.

Many of the key applications of Szemerédi's regularity lemma use an associated counting lemma, which shows that the count of every small subgraph across parts is close to what is expected if the edges go across pairs which are regular. In order to prove extensions of applications of the regularity method which applies to sparse graphs, it remained a well-known open problem to prove a counting lemma in sparse graphs. For general graphs, counterexamples are known. In random graphs, proving such an embedding lemma is a famous problem, known as the KŁR conjecture [79], which has only been resolved very recently [6, 29, 110].

Establishing an analogous result in pseudorandom graphs has been a central problem in this area. Roughly, a graph is pseudorandom if it satisfies certain properties that a random graph of the same density typically satisfies. Certain partial results are known in this case [81, 82], but it has remained an open problem to prove a counting lemma for embedding a general fixed subgraph. This problem was recently resolved by Conlon, Zhao, and the author [26]. As a consequence, many classical results in extremal combinatorics are extended to sparse pseudorandom graphs.

For example, we have the following sparse graph removal lemma. The following is a standard notion of pseudorandomness in graphs. A graph Γ is (p, β)-*jumbled* if for all vertex subsets X, Y, we have $|e(X, Y) - p|X||Y|| \leq \beta\sqrt{|X||Y|}$.

Theorem 5.1. *[26] For each $\epsilon > 0$, there is $c > 0$ and $\delta > 0$ such that if $\beta \leq cp^{t+1}n$, then any (p, β)-jumbled graph Γ has the following property. Any subgraph of Γ containing at most $\delta p^{\binom{t}{2}} n^t$ copies of K_t can be made K_t-free by removing at most ϵpn^2 edges.*

The celebrated Green-Tao theorem [69] states that the primes contain arbitrarily long arithmetic progressions. The proof of the Green-Tao theorem has two steps. The first step is establishing a *relative Szemerédi theorem*, which states that any relatively dense subset of a pseudorandom set of integers must contain arbitrarily long arithmetic progressions. The second part is finding a pseudorandom set of "almost primes" which contains, but is not much larger than, the primes. The research on the sparse counting lemma led Conlon, Zhao,

and the author [27] to prove a new relative Szemerédi theorem that requires a substantially weaker pseudorandomness condition. The proof develops the regularity method for sparse hypergraphs. This simplifies the proof of the Green-Tao theorem, see [28] for a recent exposition of the proof.

5.5. Arithmetic regularity lemma. Green [68] proved an arithmetic regularity lemma, and deduced the following arithmetic removal lemma. For each $\epsilon > 0$ and integer $m \geq 3$, there is $\delta > 0$ such that if G is an abelian group of order N, and $A_1, \ldots, A_m$ are subsets of G such that there are at most δN^{m-1} solutions to $a_1 + \cdots + a_m = 0$ with $a_i \in A_i$ for all i, then it is possible to remove at most ϵN elements from each set A_i so as to obtain sets A_i' for which there are no solutions to $a_1' + \cdots + a_m' = 0$ with $a_i' \in A_i'$ for all i. Král, Serra, and Vena [86] found a simple proof of this arithmetic removal lemma using the graph removal lemma which extends to all groups. The author's improvement on the bound in the graph removal lemma yields a similar improvement for the arithmetic removal lemma (see [47]).

Using the hypergraph regularity method, Shapira [112] and independently Král, Serra, and Vena [86] proved a conjecture of Green establishing a removal lemma for systems of linear equations. Green and Tao [71] develop an arithmetic regularity method based on the Gowers uniformity norm and deduce numerous consequences.

5.6. Semi-algebraic regularity lemma. A *k-uniform semi-algebraic hypergraph* $H=(V, E)$ consists of a vertex set of points $V \subset \mathbb{R}^d$ and an edge set E consisting of those k-tuples of points in V which satisfy a particular Boolean combination of finite polynomial equations and inequalities in kd real variables. The description complexity is the maximum of k, d, the degrees of the polynomial equations and inequalities, and the number of these polynomial equations and inequalities. Alon, Pach, Pinchasi, Radoičič, and Sharir [2] proved a regularity lemma for semi-algebraic graphs ($k = 2$). This was later extended to hypergraphs by Gromov, Lafforgue, Naor, Pach, and the author [49]. This results states that for each D and $\epsilon > 0$, there is $L = L(\epsilon, D)$ such that every k-uniform semi-algebraic hypergraph $H = (V, E)$ of description complexity at most D has an equitable vertex partition $V = V_1 \cup \ldots \cup V_\ell$ into $\ell \leq L$ parts so that all but an ϵ-fraction of the k-tuples $V_{i_1} \times \cdots \times V_{i_k}$ are complete or empty, i.e. is a subset of E or disjoint from E. Recently, Pach, Suk, and the author [54] proved that we may take the number of parts L in the semi-algebraic hypergraph regularity lemma to be polynomial in $1/\epsilon$. These results have numerous consequences in discrete geometry.

Acknowledgements. The author is supported by a Packard Fellowship, by NSF CAREER award DMS-1352121, by an Alfred P. Sloan Fellowship, and by an MIT NEC Corporation Award. The author would also like to thank David Conlon for many helpful comments.

References

[1] N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy, *Efficient testing of large graphs*, Combinatorica **20** (2000), 451–476.

[2] N. Alon, J. Pach, R. Pinchasi, R. Radoičić, and M. Sharir, *Crossing patterns of semi-algebraic sets*, J. Combin. Theory Ser. A **111** (2005), 310–326.

[3] N. Alon and J. H. Spencer, *The probabilistic method*, 3rd ed., Wiley, 2008.

[4] N. Alon and A. Shapira, *A characterization of the (natural) graph properties testable with onesided error*, in SIAM J. Comput. (Special Issue on FOCS '05) **37** (2008), 1703–1727.

[5] F. V Atkinson, G. A. Watterson, and P. A. D. Moran, *A matrix inequality*, Quart. J. Math. Oxford II **11** (1960), 137–140.

[6] J. Balogh, R. Morris, and W. Samotij, *Independent sets in hypergraphs*, preprint.

[7] I. Benjamini and Y. Peres, *A correlation inequality for tree-indexed Markov chains*, in Seminar on Stochastic Processes, Proc. Semin., Los Angeles/CA (USA) 1991, Prog. Probab. 29, 1992, 7–14.

[8] G. R. Blakley and P. A. Roy, *A Hölder type inequality for symmetric matrices with nonnegative entries*, Proc. Amer. Math. Soc. **16** (1965) 1244–1245.

[9] B. Bollobás, *Random graphs* (2nd edition), Cambridge Studies in Advanced Mathematics 73, Cambridge Univ. Press, 2001.

[10] B. Bollobás and P. Erdős, *On a Ramsey-Turán type problem*, J. Combin. Theory Ser. B. **21** (1976) 166–168.

[11] S.A. Burr and P. Erdős, *On the magnitude of generalized Ramsey numbers for graphs*, in Infinite and Finite Sets, Vol. 1 (Keszthely, 1973), Colloq. Math. Soc. János Bolyai, Vol. 10, North-Holland, Amsterdam/London, 1975, 215–240.

[12] P. Catlin, *Hajós' graph-coloring conjecture: variations and counterexamples*, J. Combin. Theory Ser. B **26** (1979), 268–274.

[13] M. Chudnovsky, *The Erdös-Hajnal conjecture – a survey*, J. Graph Theory **75** (2014), 178–190.

[14] F. R. K. Chung and R. L. Graham, *Quasi-random set systems*, J. Amer. Math. Soc. **4** (1991), 151–196.

[15] _____, *Quasi-random graphs with given degree sequences*, Random Structures Algorithms **12** (2008), 1–19.

[16] F. R. K. Chung, R. L. Graham, and R. M. Wilson, *Quasi-random graphs*, Combinatorica **9** (1989), 345–362.

[17] V. Chvatál, V. Rödl, E. Szemerédi, and W. T. Trotter Jr., *The Ramsey number of a graph with bounded maximum degree*, J. Combin. Theory Ser. B **34** (1983), 239–243.

[18] D. Conlon, *A new upper bound for diagonal Ramsey numbers*, Ann. of Math. **170** (2009), 941–960.

[19] _____, *Hypergraph packing and sparse bipartite Ramsey numbers*, Combin. Probab. Comput. **18** (2009), 913–923.

[20] D. Conlon and J. Fox, *Bounds for graph regularity and removal lemmas*, Geom. Funct. Anal. **22** (2012), 1191–1256.

[21] ———, *Graph removal lemmas*, Surveys in Combinatorics, Cambridge University Press, 2013, 1–50.

[22] D. Conlon, J. Fox, and B. Sudakov, *An approximate version of Sidorenko's conjecture*, Geom. Funct. Anal. **20** (2010), 1354–1366.

[23] ———, *Ramsey numbers of sparse hypergraphs*, Random Structures Algorithms **35** (2009), 1–14.

[24] ———, *On two problems in graph Ramsey theory*, Combinatorica **32** (2012), 513–535.

[25] ———, *Two extensions of Ramsey's theorem*, Duke Math. J. **162** (2013) 2903–2927.

[26] D. Conlon, J. Fox, and Y. Zhao, *Extremal results in sparse pseudorandom graphs*, Adv. Math. **256** (2014), 206–290.

[27] ———, *A relative Szemerédi theorem*, preprint.

[28] ———, *The Green-Tao theorem: an exposition*, preprint.

[29] D. Conlon, W. T. Gowers, W. Samotij, and M. Schacht, *On the KŁR conjecture in random graphs*, preprint.

[30] O. Cooley, N. Fountoulakis, D. Kühn, and D. Osthus, *3-uniform hypergraphs of bounded degree have linear Ramsey numbers*, J. Combin. Theory Ser. B **98** (2008), 484–505.

[31] ———, *Embeddings and Ramsey numbers of sparse k-uniform hypergraphs*, Combinatorica **28** (2009), 263–297.

[32] D. Dellamonica, S. Kalyanasundaram, D. Martin, V. Rödl and A. Shapira, *An optimal algorithm for computing Frieze-Kannan regular partitions*, Combin. Probab. and Comput., to appear.

[33] W. Deuber, *A generalization of Ramsey's theorem*, in Infinite and Finite Sets, Vol. 1 (Keszthely, 1973), Colloq. Math. Soc. János Bolyai, Vol. 10, North-Holland, Amsterdam/London, 1975, 323– 332.

[34] G. Dirac, *A property of 4-chromatic graphs and some remarks on critical graphs*, J. London Math. Soc. **27** (1952), 85–92.

[35] R. A. Duke, H. Lefmann, and V. Rödl, *A fast approximation algorithm for computing the frequencies of subgraphs in a given graph*, SIAM J. Comput. **24** (1995), 598–620.

[36] N. Eaton, *Ramsey numbers for sparse graphs*, Discrete Math. **185** (1998), 63–75.

[37] P. Erdős, *Some remarks on the theory of graphs*, Bull. Amer. Math. Soc. **53** (1947), 292–294.

[38] ———, *Problems and results on finite and infinite graphs*, in Recent advances in graph theory (Proc. Second Czechoslovak Sympos., Prague, 1974), Academia, Prague, 1975, 183–192.

[39] ———, *On the combinatorial problems which I would most like to see solved*, Combinatorica **1** (1981), 25–42.

[40] P. Erdős and S. Fajtlowicz, *On the conjecture of Hajós*, Combinatorica **1** (1981), 141–143.

[41] P. Erdős and A. Hajnal, *Ramsey-type theorems*, Discrete Appl. Math. **25** (1989), 37–52.

[42] P. Erdős, A. Hajnal, and J. Pach, *On a metric generalization of Ramsey's theorem*, Israel J. Math. **102** (1997), 283–295.

[43] _____, *Ramsey-type theorem for bipartite graphs*, Geombinatorics **10** (2000), 64–68.

[44] P. Erdős, A. Hajnal, and L. Pósa, *Strong embeddings of graphs into colored graphs*, in Infinite and Finite Sets, Vol. 1 (Keszthely, 1973), Colloq. Math. Soc. János Bolyai, Vol. 10, North-Holland, Amsterdam/London, 1975, 585–595.

[45] P. Erdős and G. Szekeres, *A combinatorial problem in geometry*, Composito Math. **2** (1935), 463–470.

[46] P. Erdős and E. Szemerédi, *On a Ramsey type theorem*, Period. Math. Hungar. **2** (1972) 295–299.

[47] J. Fox, *A new proof of the graph removal lemma*, Ann. of Math. **174** (2011), 561–579.

[48] J Fox, A. Grinshpun, L. M. Lovász, and Y. Zhao, *On regularity lemmas and their applications*, in preparation.

[49] J. Fox, M. Gromov, V. Lafforgue, A. Naor, and J. Pach, *Overlap properties of geometric expanders*, J. Reine Angew. Math. **671** (2012), 49–83.

[50] J. Fox, C. Lee, and B. Sudakov, *Chromatic number, clique subdivisions, and the conjectures of Hajós and Erdős-Fajtlowicz*, Combinatorica **33** (2013), 181–197.

[51] J. Fox, and P. Loh, *On a problem of Erdős and Rothschild on edges in triangles*, Combinatorica **32** (2012), 619–628.

[52] J. Fox, P. Loh, and Y. Zhao, *The critical window for the classical Ramsey-Turán problem*, Combinatorica, to appear.

[53] J. Fox and L. M. Lovász, *A tight lower bound for Szemerédi's regularity lemma*, preprint.

[54] J. Fox, J. Pach, and A. Suk, *Density and regularity theorems for semi-algebraic hypergraphs*, preprint.

[55] J. Fox and B. Sudakov, *Induced Ramsey-type theorems*, Adv. Math. **219** (2008), 1771–1800.

[56] _____, *Density theorems for bipartite graphs and related Ramsey-type results*, Combinatorica **29** (2009), 153–196.

[57] _____, *Two remarks on the Burr-Erdős conjecture*, European J. Combin. **30** (2009), 1630–1645.

[58] _____, *Dependent random choice*, Random Structures Algorithms **38** (2011), 68–99.

[59] A. Frieze and R. Kannan, *Quick approximation to matrices and applications*, Combinatorica **19** (1999), 175–220.

[60] S. Gerke and A. Steger, *The sparse regularity lemma and its applications*, In Surveys in Combinatorics 2005, volume 327 of London Math. Soc. Lecture Note Ser., pp. 227–258. Cambridge Univ. Press, Cambridge, 2005.

[61] W. T. Gowers, *Lower bounds of tower type for Szemerédi's uniformity lemma*, Geom. Funct. Anal. **7** (1997), 322–337.

[62] ———, *A new proof of Szemerédi's theorem for arithmetic progressions of length four*, Geom. Funct. Anal. **8** (1998), 529–551.

[63] ———, *A new proof of Szemerédi's theorem*, Geom. Funct. Anal. **11** (2001), 465–588.

[64] ———, *Hypergraph regularity and the multidimensional Szemerédi theorem*, Ann. of Math. **166** (2007), 897–946.

[65] ———, *Quasirandom groups*, Combin. Probab. Comput. **17** (2008), 363–387.

[66] R. L. Graham, V. Rödl, and A. Ruciński, *On graphs with linear Ramsey numbers*, J. Graph Theory 35 (2000), 176–192.

[67] ———, *On bipartite graphs with linear Ramsey numbers*, Combinatorica **21** (2001), 199–209.

[68] B. Green, *A Szemerédi-type regularity lemma in abelian groups, with applications*, Geom. Funct. Anal. **15** (2005), 340–376.

[69] B. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions*, Ann. of Math. **167** (2008), 481–547.

[70] ———, *Linear equations in primes*, Ann. of Math. **171** (2010), 1753–1850.

[71] ———, *An arithmetic regularity lemma, an associated counting lemma, and applications*, An irregular mind, Bolyai Soc. Math. Stud., vol. 21, János Bolyai Math. Soc., Budapest, 2010, 261–334.

[72] ———, *The Möbius function is strongly orthogonal to nilsequences*, Ann. of Math. 175 (2012), 541–566.

[73] ———, *The quantitative behaviour of polynomial orbits on nilmanifolds*, Ann. of Math. **175** (2012), 465–540.

[74] B. Green, T. Tao, and T. Ziegler, *An inverse theorem for the Gowers $U^{s+1}[N]$-norm*, Ann. of Math. **176** (2012), 1231–1372.

[75] H. Hatami, *Graph norms and Sidorenko's conjecture*, Israel J. Math. **175** (2010) 125–150.

[76] S. Kalyanasundaram and A. Shapira, *A wowzer-type lower bound for the strong regularity lemma*, Proc. London Math. Soc. **106** (2013), 621–649.

[77] J. H. Kim, C. Lee, and J. Lee, *Two approaches to Sidorenko's conjecture*, preprint.

[78] Y. Kohayakawa, *Szemerédi's regularity lemma for sparse graphs*, In Foundations of computational mathematics (Rio de Janeiro, 1997), pp. 216–230. Springer, Berlin, 1997.

[79] Y. Kohayakawa, T. Łuczak, and V. Rödl, *On K_4-free subgraphs of random graphs*, Combinatorica **17** (1997), 173–213.

[80] Y. Kohayakawa, H. Prömel, and V. Rödl, *Induced Ramsey numbers*, Combinatorica **18** (1998), 373–404.

[81] Y. Kohayakawa, V. Rödl, M. Schacht, and J. Skokan, *On the triangle removal lemma for subgraphs of sparse pseudorandom graphs*, In An Irregular Mind (Szemerédi is 70), volume 21 of Bolyai Soc. Math. Stud., pp. 359–404. Springer Berlin, 2010.

[82] Y. Kohayakawa, V. Rödl, and P. Sissokho, *Embedding graphs with bounded degree in sparse pseudorandom graphs*, Israel J. Math. **139** (2004), 93–137.

[83] J. Komlós and M. Simonovits, *Szemerédi's regularity lemma and its applications in graph theory*, Combinatorics, in Paul Erdős is eighty, Vol. 2 (Keszthely, 1993), 295–352, Bolyai Soc. Math. Stud., 2, János Bolyai Math. Soc., Budapest, 1996.

[84] A. Kostochka and V. Rödl, *On graphs with small Ramsey numbers*, J. Graph Theory **37** (2001), 198–204.

[85] A. Kostochka and B. Sudakov, *On Ramsey numbers of sparse graphs*, Combin. Probab. Comput. **12** (2003), 627–641.

[86] D. Král, O. Serra, and L. Vena, *A combinatorial proof of the removal lemma for groups*, J. Combin. Theory Ser. A **116** (2009), 971–978.

[87] ———, *A removal lemma for systems of linear equations over finite fields*, Israel J. Math. **187** (2012), 193–207.

[88] M. Krivelevich and B. Sudakov, *Pseudorandom graphs*, in More Sets, Graphs and Numbers, Bolyai Society Mathematical Studies 15, Springer, 2006, 199–262.

[89] J. X. Li and B. Szegedy, *On the logarithimic calculus and Sidorenko's conjecture*, preprint.

[90] L. Lovász, *Very large graphs*, Current Developments in Mathematics Volume 2008 (2009), 67–128.

[91] L. Lovász, *Large networks and graph limits*, American Mathematical Society Colloquium Publications, 60. American Mathematical Society, Providence, RI, 2012.

[92] L Lovász and B. Szegedy, *Szemerédi's lemma for the analyst*, Geom. Funct. Anal. **17** (2007), 252–270.

[93] T. Łuczak and V. Rödl, *On induced Ramsey numbers for graphs with bounded maximum degree*, J. Combin. Theory Ser. B **66** (1996), 324–333.

[94] G. Moshkovitz and A. Shapira, *A short proof of Gowers' lower bound for the regularity lemma*, preprint.

[95] B. Nagle, S. Olsen, V. Rödl, and M. Schacht, *On the Ramsey number of sparse 3-graphs*, Graphs Combin. **27** (2008), 205–228.

[96] B. Nagle, V. Rödl, and M. Schacht, *The counting lemma for regular k-uniform hypergraphs*, Random Structures Algorithms **28** (2006), 113–179.

[97] J. Nešetřil and J. A. Väänänen, *Combinatorics and quantifiers*, Comment. Math. Univ. Carolin. **37** (1996), 433–443.

[98] R. Pemantle and Y. Peres, *Domination between trees and application to an explosion problem*, Ann. Probab. **22** (1994), 180–194.

[99] Y. Peng, V. Rödl, and A. Ruciński, *Holes in graphs*, Electron. J. Combin. **9** (2002), Research Paper 1, 18 pp. (electronic).

[100] H. Prömel and V. Rödl, *Non-Ramsey graphs are $c \log n$-universal*, J. Combin. Theory Ser. A **88** (1999) 379–384.

[101] F. P. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. **30** (1930), 264–286.

[102] V. Rödl, *The dimension of a graph and generalized Ramsey theorems, Master's thesis, Charles University*, 1973.

[103] ———, *On universality of graphs with uniformly distributed edges*, Discrete Math. **59** (1986), 125–134.

[104] ———, *On homogeneous sets of positive integers*, J. Combin. Theory Ser. A **102** (2003), 229–240.

[105] ———, *Quasi-randomness and the regularity method in hypergraphs*, Proc. of the International Congress of Mathematicians (ICM), Seoul 2014, Korea, to appear.

[106] V. Rödl and M. Schacht, *Regularity lemmas for graphs*, Fete of Combinatorics and Computer Science, vol. 20 series (2010) Bolyai Soc. Math. Stud., 287–325.

[107] V. Rödl and J. Skokan, *Regularity lemma for uniform hypergraphs*, Random Structures Algorithms **25** (2004), 1–42.

[108] K. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 104–109.

[109] I. Z. Ruzsa and E. Szemerédi, *Triple systems with no six points carrying three triangles*, in Combinatorics (Keszthely, 1976), Coll. Math. Soc. J. Bolyai 18, Volume II, 939–945.

[110] D. Saxton and A. Thomason, *Hypergraph containers*, preprint.

[111] A. Scott, *Szemerédi's regularity lemma for matrices and sparse graphs*, Combin. Probab. Comput. **20** (2011), 455–466.

[112] A. Shapira, *A proof of Green's conjecture regarding the removal properties of sets of linear equations*, J. London Math. Soc. **81** (2010), 355–373.

[113] S. Shelah, *A finite partition theorem with double exponential bound*, The mathematics of Paul Erdős, II, 240–246, Algorithms Combin., 14, Springer, Berlin, 1997.

[114] A. F. Sidorenko, *A correlation inequality for bipartite graphs*, Graphs Combin. **9** (1993), 201–204.

[115] _____, *An analytic approach to extremal problems for graphs and hypergraphs*, in Extremal problems for finite sets (Visegrád, 1991), 423–455, Bolyai Soc. Math. Stud., 3, János Bolyai Math. Soc., Budapest, 1994.

[116] M. Simonovits, *Extremal graph problems, degenerate extremal problems and supersaturated graphs*, in Progress in graph theory (Waterloo, Ont., 1982), Academic Press, Toronto, ON, 1984, 419–437.

[117] M. Simonovits and V.T. Sós, *Ramsey-Turán theory*, Discrete Math. **229** (2001), 293–340.

[118] J. Skokan and L. Thoma, *Bipartite subgraphs and quasi-randomness*, Graphs Combin. **20** (2004), 255–262.

[119] V. T. Sós, *On extremal problems in graph theory, in Proceedings of the Calgary International Conference on Combinatorial Structures and their Application*, 1969, 407–410.

[120] B. Sudakov, *A few remarks on the Ramsey-Turán-type problems*, J. Combin. Theory Ser. B **88** (2003), 99–106.

[121] E. Szemerédi, *On graphs containing no complete subgraph with 4 vertices (Hungarian)*, Mat. Lapok **23** (1972) 113–116.

[122] _____, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 199–245.

[123] _____, *Regular partitions of graphs*, in Colloques Internationaux CNRS 260 - Problémes Combinatoires et Théorie des Graphes, Orsay (1976), 399–401.

[124] T. Tao, *Higher order Fourier analysis*, Graduate Studies in Mathematics 142, American Mathematical Society, 2012.

[125] _____, *Szemerédi's regularity lemma revisited*, Contrib. Discrete Math. **1** (2006), 8–28.

[126] A. G. Thomason, *Pseudorandom graphs*, in Random graphs '85 (Poznań, 1985), North-Holland Math. Stud., vol. 144, North-Holland, Amsterdam, 1987, 307–331.

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA

E-mail: fox@math.mit.edu

Positional games

Michael Krivelevich

Abstract. Positional games are a branch of combinatorics, researching a variety of two-player games, ranging from popular recreational games such as Tic-Tac-Toe and Hex, to purely abstract games played on graphs and hypergraphs. It is closely connected to many other combinatorial disciplines such as Ramsey theory, extremal graph and set theory, probabilistic combinatorics, and to computer science. We survey the basic notions of the field, its approaches and tools, as well as numerous recent advances, standing open problems and promising research directions.

Mathematics Subject Classification (2010). Primary 05C57, 91A46; Secondary 05C80, 05D05, 05D10, 05D40.

Keywords. Positional games, Ramsey theory, extremal set theory, probabilistic intuition.

1. Introductory words

Positional games are a combinatorial discipline, whose basic aim is to provide a mathematical foundation for analysis of two-player games, ranging from popular recreational games such as Tic-Tac-Toe and Hex to purely abstract games played on graphs and hypergraphs. Though the field has been in existence for several decades, motivated partly by its recreational side, it advanced tremendously in the last few years, maturing into one of the central branches of modern combinatorics. It has been enjoying mutual and fruitful interconnections with other combinatorial disciplines such as Ramsey theory, extremal graph and set theory, probabilistic combinatorics, as well as theoretical computer science.

The aim of this survey is two-fold. It is meant to provide a brief, yet gentle, introduction to the subject to those with genuine interest and basic knowledge in combinatorics. At the same time, we cover recent progress in the field, as well as its standing challenges and open problems. We also put a special emphasis on connections between positional games and other branches of combinatorics, in particular discussing the very surprising ubiquitous role of probabilistic intuition and considerations in the analysis of (entirely deterministic) positional games.

Due to obvious space limitations we will frequently be rather brief, omitting many of the proofs or merely indicating their outlines. More details, examples and discussions can be found in research monographs and papers on the subject.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

2. Basic setting and examples

Positional games involve two players alternately claiming unoccupied elements of a set X, the *board* of the game; the elements of X are called vertices. Usually X is assumed to be finite, although of course there are exciting infinite games to analyze. The focus of players' attention is a given family $\mathcal{H} = \{A_1, \ldots, A_k\}$ of subsets of X, called the *hypergraph of the game*; sometimes the members of $\mathcal{H}$ are referred to as the *winning sets* of the game. In the most general version there are two additional parameters — positive integers p and q: the first player claims p unoccupied elements in each turn, the second player answers by claiming q vertices. (If in the very end of the game there are less unclaimed elements to claim than as prescribed by the turn of the current player, that player claims all remaining elements.) The parameters p and q define the *bias* of the game. The most basic case $p = q = 1$ is the so called *unbiased* game. The game is specified completely by setting its outcome (first player's win/second player's win/draw) for every final position of the game, or more generally for every possible game scenario (an alternating sequence of legal moves of both players). For every game scenario there is only one possible outcome. Of course, the above definition is utterly incomplete and hence fairly vague. However, the accumulated research experience has shown that this is the right setting for the field. Depending on concrete game rules we get several game types, some of which are discussed later. For now let us state, using the standard game theory terminology, that positional games are two-player perfect information zero sum games with no chance moves.

Now we illustrate the above general setting by providing several examples.

Example 2.1 (Tic-Tac-Toe). This is of course the first game that should come to anyone's mind. In our terminology, the board of the game X is the 3-by-3 square. Two players, sometimes called Crosses and Noughts, claim in their turns one unoccupied element of the board each. The winning sets are three horizonal lines, three vertical lines, and two diagonals, all of size three; thus the game hypergraph $\mathcal{H}$ has eight sets of size three each, and is thus 3-uniform. (A hypergraph is called *r-uniform* if all its edges are of size r). The player completing a winning set first wins; if none of the lines is claimed by either player in the end, the game is declared a draw. Assuming optimal strategies of both players, the game is a draw, as everybody knows; here the case analysis is essentially the only way to prove it.

Example 2.2 (n^d). This is a natural, yet extremely far reaching and challenging generalization of the classical Tic-Tac-Toe game. Given positive integers d and n, the board X of n^d is the d-dimensional cube $X = [n]^d$, and the winning sets are the so-called *geometric lines* in X. A geometric line l is a family of n distinct points $\left(\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(n)}\right)$ of X, where $\mathbf{a}^{(i)} = (a_1^{(i)}, \ldots, a_d^{(i)})$, such that for each $1 \leq j \leq d$ the sequence of corresponding coordinates $(a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(n)})$ is either $(1, 2, \ldots, n)$, or $(n, n-1, \ldots, 1)$, or constant (and of course at least one of the coordinates should be non-constant). The winner is the player who occupies a whole line first, otherwise the game ends in a draw. The familiar Tic-Tac-Toe is 3^2 in this notation. Another recreationally known instance of this game is 4^3, marketed by Parker Brothers under the name of Qubic.

This is a very complicated game, and resolving it for all pairs (d, n) appears to be well out of reach. We can do it for the fairly simple two-dimensional case $d = 2$ (first player's win for $n = 1, 2$, and a draw for $n \geq 3$), and also for the two extremes: n large compared to d, and d large compared to n. For the former case, we have the following:

Theorem 2.3. *If $n \geq 3^d - 1$, then the n^d game is a draw.*

For the proof, one can argue that the maximum vertex degree in the game hypergraph of n^d is $(3^d - 1)/2$ for odd n and is $2^d - 1$ for even n; then a straightforward application of Hall's theorem shows that one can assign to every geometric line l a pair of points $B(l) \subset l$ so that the assigned pairs are disjoint. Given such an assignment, either player can guarantee that he will not lose by claiming at least one element from each pair. This is a good example of the so called *pairing strategy* argument, frequently used to show the draw outcome in games of this type.

The case of fixed n and large d is first player's win. This is the famous Hales-Jewett theorem, and we will discuss it soon.

Example 2.4 (Sim). This too is a well known recreational game, whose mathematical description is as follows. The board of the game is the edge set of the complete graph K_6 on six vertices. Two players claim alternately one unoccupied edge of the board each, and the player completing a triangle of his edges first actually loses. Thus the game hypergraph $\mathcal{H}$ consists of the edge sets of all triangles in K_6, so $\mathcal{H}$ has $\binom{6}{3} = 15$ sets of size 3 each. This is a reverse, or misére-type, game, the player completing a winning set first is the one to lose. Due to the standard fact $R(3,3) = 6$, where $R(k,l)$ is the Ramsey number, the game cannot end in a draw. Sim has been solved by a computer and was proven to be a second's player win, with a complicated winning strategy.

Example 2.5 (Hex). This game was invented by the Danish scientist Piet Hein in 1942 and was played and researched by John Nash in his student years. It is played on a rhombus of hexagons of size $n \times n$ (in recreational versions n is usually 11), where two players, say, Blue and Red, take the two opposite sides of the board each, and then alternately mark unoccupied hexagons of the board with their own color. Whoever connects his own opposite sides of the board first wins the game.

There is a catch here — the game as described above does not fit our general scheme, as the players' winning sets are different. The problem can be cured by proving that a player wins in Hex if and only if he prevents his opponent from winning, by blocking his winning ways. This is the so-called Hex theorem, known to be equivalent to the Brouwer fixed point theorem, see [37]. This allows to cast Hex in our general framework, by defining the winning sets to be the connecting sets of hexagons for the first player and declaring him the winner if he occupies one of them fully in the end; the second player's goal is redefined by assigning him instead the task of preventing the first player from claiming an entire winning set.

Example 2.6 (Connectivity game). The game is played on the edge set of a multigraph G. The players, called Connector and Disconnector, take turns in claiming one unoccupied edge of G each, with Disconnector moving first. Connector wins the game if in the end the set of his edges contains a spanning tree of G, Disconnector wins if he manages to leave Connector with a non-connected graph. Observe the highly non-symmetric goals of the players here. In our setting, the board is the set of edges of G, the winning sets are the edge sets of spanning trees in G; Connector wins the game if in the end he claims fully one of the winning sets, Disconnector wins otherwise. This game was treated by Lehman, who proved:

Theorem 2.7 ([60]). *If a multigraph G has two edge-disjoint spanning trees, then Connector wins the connectivity game played on G.*

The proof is by induction on $|V(G)|$. For the induction step, assume that T_1 and T_2 are edge-disjoint spanning trees of G. If Disconnector claims an edge e from, say, T_1, this move cuts

T_1 into two connected parts. Connector responds by claiming an edge f of T_2 connecting these two parts, and then contracts f by identifying its two endpoints u and v. Then he updates T_1 and T_2 accordingly and applies induction to G'. The case where Disconnector claims an edge outside of $T_1 \cup T_2$ can be treated similarly.

Frequently the board of the game is the edge set of the complete graph K_n, and the players take turns in occupying its edges. Here is one such example.

Example 2.8 (Hamiltonicity game). This game is set up as follows. It is played by two players, who take turns claiming unoccupied edges of K_n, one edge each turn. The first player wins if by the end of the game he manages to make a Hamilton cycle (a cycle of length n) from his edges, while the second player wins otherwise, i.e., if in the end he manages to put his edge, or to break, into every Hamilton cycle. Here, the board is $E(K_n)$, and the winning sets are n-cycles; the first player wins if he claims an entire winning set by the end of the game, and the second player wins otherwise. This game was introduced and analyzed by Chvátal and Erdős in their seminal paper [19]; it turned out to be an easy win for the cycle maker for all large enough n.

Example 2.9 (Row-column game). The board of the game is the $n \times n$ square, and two players alternately claim its elements. The goal of the first player is to achieve a sizable advantage in some row or column; the question is how large an advantage he can get playing against a perfect opponent. We can place the game in our general framework as follows. For a given parameter k, if the task is to decide whether the first player can reach at least k elements in one of the $2n$ lines of the game, one can define the game hypergraph whose board consists of the n^2 cells of the square, and whose winning sets are all k-subsets of the rows and columns.

If only rows (or only columns) are taken into account, then a simple pairing strategy shows that the first player can achieve nothing for even n, or 1 for odd n. However, when both rows and columns are important, the first player can reach something of substance: Beck proved [9] that he has a strategy to end up with at least $n/2 + 32\sqrt{n}$ elements in some line. The upper bound is due to Székely, who showed [71] that the second player can restrict his opponent to at most $n/2 + O(\sqrt{n \log n})$ elements in each line; the gap of $c\sqrt{\log n}$ in the error term still stands.

There is a crucial difference between casual games, where more experienced or skillful players have better chances to succeed, and formal games we consider here. We assume that both players have unbounded computational power and therefore play perfectly, choosing optimal moves each turn. Under this assumption, each positional game is *determined* and has exactly one of the three possible mutually exclusive outcomes: the first player has a winning strategy, the second player has a winning strategy, or both players have drawing strategies. A formal proof of this statement is easy and uses De Morgan's laws. Thus, solving a game means establishing its deterministic outcome out of the set of three possibilities. In order to do it formally, one can employ the fairly natural notion of a *game tree*. Given a positional game, its game tree is a rooted directed tree, where each node corresponds to a sequence of legal moves of both players, including the empty sequence — the root of the tree, and there is an incoming edge to every legal sequence of moves from the sequence one move shorter. The leaves are exactly the final positions of the game. In order to find the outcome of the game, one can backtrack the game tree by labeling first its roots by the corresponding game result, and then for each intermediate node marking it with the best possible move out of this position. Although this simple procedure resolves every positional game, in reality it is

usually extremely impractical, due to the huge size of the game tree. To illustrate this thesis, let us mention that the 4^3 game, or the Qubic, is known to be first player's win, but according to Oren Patashnik, one of its solvers, the winning strategy fully spelled would be as thick as a phone book. This leaves us — luckily, in fact — with the necessity to develop general combinatorial tools for analyzing positional games.

Let us say a couple of words on where it all belongs mathematically. The term "positional games" can be somewhat misleading. Classical game theory is largely based on the notions of uncertainty and lack of perfect information, giving rise to probabilistic arguments and the key concept of a mixed strategy. Positional games, in contrast, are perfect information games and as such can in principle be solved completely by an all-powerful computer and hence are categorized as trivial in classical game theory. This is not the case of course, due to the prohibitive complexity of the exhaustive search approach; this only stresses the importance of accessible mathematical criteria for analyzing such games. A probably closer relative is what is sometimes called "combinatorial game theory", popularized by Conway and others, which includes such games as Nim; they are frequently based on algebraic arguments and notions of decomposition. Positional games are usually quite different and call for combinatorial arguments of various sorts.

Now we wish to mention a few papers and researchers who were influential in the development of the field. Of course, the choice below is rather subjective. The paper of Hales and Jewett [43] has established a tight connection between positional games and Ramsey theory. As is the case with many combinatorial disciplines, Paul Erdős was a pioneer here. First he wrote the paper [27] with Selfridge, where potential functions were used for game analysis, at the same time providing the first derandomization argument, a crucial notion in the theory of algorithms. Then Chvátal and Erdős [19] introduced biased games. József Beck shaped the field for the last three decades with his many papers that were instrumental in turning positional games into a coherent combinatorial discipline.

We complete this introductory section with suggestions for further reading. The fundamental monograph [11] of Beck covers many facets of positional games. His more recent book [12] contains a lot of material on games too. The new book [49] can serve as a gentle introduction to the subject, at the same time covering recent important developments.

3. Strong games

Strong games are probably the most natural, at least from the layman perspective, type of games. A *strong game* is played on a game hypergraph $(X, \mathcal{H})$ by two players, called First Player or FP, and Second Player, or SP, who take turns in occupying previously unclaimed elements of X, one element each time; First Player starts. The winner is the *first* player to occupy completely a winning set $A \in \mathcal{H}$; if this has not happened till the end of the game, it is declared a draw. Tic-Tac-Toe and n^d are games of this type.

The childhood intuition suggesting it is beneficial to be the player to move first has a solid mathematical basis:

Theorem 3.1. *In a strong game played on $(X, \mathcal{H})$, First Player can guarantee at least a draw.*

Proof. The proof applies the so-called *strategy stealing principle*, observed by Nash. Assume to the contrary that Second Player has a winning strategy $\mathcal{S}$. The strategy is a complete

recipe, prescribing SP how to respond to each move of his opponent, and to reach a win eventually. Now, First Player "steals" $\mathcal{S}$ and adopts it as follows. He starts with an arbitrary move and then pretends to be Second Player, by ignoring his first move. After each move of SP, FP consults $\mathcal{S}$ and responds accordingly. If he is told to claim an element of X which is still free, he does so; if this element has been taken by him as his previously ignored arbitrary move, he takes another arbitrary move instead. Observe that an extra move can only benefit First Player. Since $\mathcal{S}$ is a winning strategy, at some point FP claims fully a winning set, even ignoring his extra move, before SP was able to do so. It follows that First Player has a winning strategy, excluding the possibility that Second Player has a winning strategy and thus providing the desired contradiction. □

Thus, in any strong game there are only two possible outcomes: First Player's win or a draw. (Both of them happen indeed for particular games.) This is perhaps the single most powerful result in positional games — it is valid for every strong game! At the same time, it is rather useless, since due to the inexplicit nature of the proof it provides no concrete directions for FP to reach at least a draw.

There are many games for which draw is impossible. This is usually a Ramsey-type statement: if for any two-coloring of X (the colors are the moves of corresponding players) there is a monochromatic winning set $A \in \mathcal{H}$, then there is no final drawing position. We have:

Corollary 3.2. *If in a strong game played on $(X, \mathcal{F})$ there is no final drawing position, then First Player has a winning strategy.*

For example, we conclude that the clique game (K_n, K_q) (the board is the edge set of the complete graph K_n, the player completing a clique K_q first wins) is First Player's win for $n \geq R(q, q)$, since by the definition of Ramsey numbers, there is no final drawing position here. Another example is the game of Hex; Nash, in a pioneering application of the strategy stealing principle, observed that due to the board symmetry strategy stealing is applicable, and the first player wins the game. Still, no clue as for how exactly the first player should play to win in these cases...

The most inspiring instance of application of this pair (Strategy Stealing, Ramsey) is probably for the n^d game. Recall that we stated that this game is a draw for n large enough compared to d (Theorem 2.3). In the opposite direction, Hales and Jewett, in one of the cornerstone papers of modern Ramsey theory [43], proved:

Theorem 3.3. *For every k and n there exists $d_0 = d_0(k, n)$ such that for every $d \geq d_0$ every k-coloring of $X = [n]^d$ contains a monochromatic geometric line.*

(See [64, 67] for simpler proofs/better bounds on d_0.) The Hales-Jewett theorem is of course a Ramsey-type result, but it implies immediately the following nice corollary, which was apparently the original motivation behind [43]:

Corollary 3.4. *For every n there exists $d_0 = d_0(n)$ such that for every $d \geq d_0$ the game n^d is First Player's win.*

To see this, simply apply Corollary 3.2 and Theorem 3.3 with $k = 2$.

Regretfully our story about strong games nearly ends here. The above two main tools (strategy stealing, Ramsey-type arguments) exhaust our set of general tools available to handle these games. In addition, strategy stealing is very inexplicit, while Ramsey-type statements frequently provide astronomic bounds. So the situation leaves a lot to be desired. At

the present state of knowledge we are unable to resolve even most basic games. The inherent difficulty in analyzing strong games can be explained partially by the fact that they are not hypergraph monotone. By this we mean the existence of examples (pretty easy ones in fact, see, e.g., Ch. 9.4 of [12]) of game hypergraphs $\mathcal{H}$ which are wins for First Player, yet one can add an extra set A to $\mathcal{H}$ to obtain a new hypergraph $\mathcal{H}'$ which is a draw; this is what Beck calls the *extra set paradox*, and it is indeed quite disturbing.

However not all is lost, and some very nice and surprising results about particular strong games have been obtained recently. We will cover them later.

4. Maker-Breaker games

We have established that in strong games it is beneficial to be the first player to move – by Theorem 3.1 he can guarantee at least a draw. If so, and perhaps thinking more practically, the second player can lower his sights and play more defensively instead, aiming to prevent his opponent from occupying fully a winning set, or putting it differently, to "break" into every winning set. This leads naturally to the very important notion of Maker-Breaker games. Given a hypergraph $(X, \mathcal{H})$, the *Maker-Breaker game* is defined as follows. There are two players, called now Maker and Breaker, taking turns in occupying one element of X in each turn. We assume that Maker moves first, unless said otherwise. Maker wins if in the end of the game he has occupied fully a winning set $A \in \mathcal{H}$, Breaker wins otherwise, i.e., if he claims at least one element in every winning set.

Maker-Breaker games have certain similarities to strong games; Maker should probably be compared to First Player, and Breaker to Second Player. Observe, for example, that if Breaker wins against Maker on $\mathcal{H}$, then Second Player draws in the corresponding strong game, using the same strategy. However, these game types are more different than similar: Maker, unlike First Player, needs to occupy a winning set eventually, and not necessarily first, in order to win; there is no draw here. Sometimes Maker-Breaker games are also called *weak games*, to contrast them with strong games.

Going back to our examples from Section 2, we can classify some of them now as Maker-Breaker games. They are the connectivity game (Connector is our Maker — we assumed him to move second, but this is a tiny detail), the Hamiltonicity game and the row-column game. Moreover, as we explained there, Hex can be put into this framework too.

We remark that the following intuitive statement is correct: if Maker wins the game played on $\mathcal{H}$ as the second player, then he also has a winning strategy as the first player; an analogous statement is of course valid for Breaker as well.

Let us now concentrate on the prospects of each of the players, and on tools available to argue for their corresponding sides. We start with Breaker. Breaker's win is closely related to the 2-colorability problem in hypergraphs (frequently also called Property B), popularized by Erdős. Observe:

Proposition 4.1. *If the Maker-Breaker game played on a hypergraph $(X, \mathcal{H})$ is Breaker's win, then $\mathcal{H}$ is 2-colorable.*

Proof. As we mentioned, Breaker's win as the second player guarantees his win as the first player as well. We can thus imagine the two-player game on $\mathcal{H}$ where both the first and the second players think of themselves as Breaker and follow Breaker's winning strategy for the corresponding player. Thus each of them comes out as a winner, meaning that in the end

each edge A of $\mathcal{H}$ will carry the marks (the colors) of both players. It follows that $\mathcal{H}$ is 2-colorable. □

It is customary nowadays to use random coloring to argue that a hypergraph is 2-colorable. It is perhaps much less standard to realize that there is a game theoretic way to look at this problem. The following statement is easily proven through the usual random argument. Let $(X, \mathcal{H})$ be a hypergraph. If $\sum_{A\in\mathcal{H}} 2^{-|A|} < 1/2$, then $\mathcal{H}$ is 2-colorable. (Color each of the vertices of X randomly and independently in red with probability 1/2 and in blue with probability 1/2; for each edge A the bad event $E_A =$"A is monochromatic" has probability $2^{-|A|+1}$, hence Prob$[\bigcup_{A\in\mathcal{H}} E_A] \leq \sum_{A\in\mathcal{H}} 2^{-|A|+1} < 1$, and thus there is a 2-coloring of $\mathcal{H}$.) Erdős and Selfridge [27] provided a game strengthening of this result:

Theorem 4.2. *Let $(X, \mathcal{H})$ be a hypergraph. If*

$$\sum_{A\in\mathcal{H}} 2^{-|A|} < 1/2\,, \tag{4.1}$$

then $\mathcal{H}$ is Breaker's win.

Proof. At any stage of the game the board X is split into three sets: the set M of vertices claimed by Maker, the set B of vertices of Breaker, and the set F of currently free vertices. Define the potential function $\Psi = \sum_{A\in\mathcal{H}:A\cap B=\emptyset} 2^{-|A\setminus M|}$. Observe that if Maker occupies at some point of the game a winning set $A \in \mathcal{H}$ fully, then $\Psi \geq 1$ at that point. Thus, for Breaker to win it is enough to maintain the value of Ψ below 1 during the game. The initial value of the potential is less than $1/2$ by the assumption of the theorem; after the first move of Maker it increases by at most the factor of 2 and is thus less than 1, a good start. Hence it is enough to prove that Breaker has a strategy to ensure that after each round (a round here is Breaker's move, followed by Maker's move) the value of Ψ does not increase. Suppose we are in the beginning of round i with partition $X = M_{i-1} \cup B_{i-1} \cup F_{i-1}$ and potential Ψ_{i-1}. Breaker's choice b_i is natural: he chooses to claim the element b_i maximizing the potential's decrease:

$$b_i = \operatorname*{argmax}_b \sum_{\substack{A\cap B_{i-1}=\emptyset \\ b\in A}} 2^{-|A\setminus M_{i-1}|}\,.$$

If Maker then claims an element m_i, the updated value Ψ_i of the potential is:

$$\Psi_i = \Psi_{i-1} - \sum_{\substack{A\cap B_{i-1}=\emptyset \\ b_i\in A}} 2^{-|A\setminus M_{i-1}|} + \sum_{\substack{A\cap B_{i-1}=\emptyset \\ m_i\in A}} 2^{-|A\setminus M_{i-1}|} - \sum_{\substack{A\cap B_{i-1}=\emptyset \\ b_i,m_i\in A}} 2^{-|A\setminus M_{i-1}|} \leq \Psi_{i-1}\,,$$

due to the choice of b_i. □

The following construction shows that criterion (4.1) is tight. Let $X = \{c\}\cup\{l_1,\ldots,l_k\}\cup\{r_1,\ldots,r_k\}$. Define $\mathcal{H} = \{A \subset X : c \in A,\ |A \cap \{l_i, r_i\}| = 1,\ i = 1,\ldots,k\}$. Maker wins the game on $\mathcal{H}$ by first taking c, and then claiming the sibling of Breaker's move (l_i for r_i, and r_i for l_i).

The proof of the Erdős-Selfridge theorem is quite simple (not a bad thing in itself!), but the result is truly remarkable for a variety of reasons. First, it provides a concrete and very useful criterion for Breaker's win. It also serves as an inspiring example of applying potential functions in positional games. Another important feature of the proof is that it supplies

a simple polynomial (in $|X| + |\mathcal{H}|$) algorithm for Breaker to win. Essentially the same argument can be used to derandomize the random 2-coloring argument given above. In fact, this was the first instance of the *method of conditional probabilities*, an important general approach to derandomizing randomized algorithms, one of the major topics in theoretical computer science, see, e.g., [3, 61]. Thus, the field of positional games reaches well beyond its immediate scope.

Here is an example of applying the Erdős-Selfridge criterion. In the Maker-Breaker version of the clique game (K_n, K_q), Maker wins if in the end of the game he has claimed a clique of size q. Let $q(n)$ be the largest q for which Maker wins the game. We can argue that $q(n) \leq 2\log_2 n$. Indeed, the game hypergraph $\mathcal{H}$ has $\binom{n}{q}$ edges and is $\binom{q}{2}$-uniform. Hence, by (4.1) if $\binom{n}{q}2^{-\binom{q}{2}} < 1/2$, Breaker wins. Solving this inequality for $q = q(n)$ gives the claimed bound. For the lower bound on $q(n)$, recall that if $n \geq R(q, q)$, then First Player wins in the corresponding strong game, and Maker can follow his footsteps. Plugging in the standard upper bound $R(q, q) < 4^q$ we derive $q(n) \geq \frac{1}{2}\log_2 n$. Beck discusses the asymptotically tight lower bound $q(n) \geq (2 - o(1))\log_2 n$ in his book [11].

For Maker's side, Beck proved [7] the following criterion.

Theorem 4.3. *Let $(X, \mathcal{H})$ be an r-uniform hypergraph. If $|\mathcal{H}| > 2^{r-3} \cdot \Delta_2(\mathcal{H}) \cdot |X|$, where $\Delta_2(\mathcal{H}) = \max\{deg(x, y) : x \neq y \in X\}$ and $deg(x, y) = |\{A \in \mathcal{H} : x, y \in A\}|$, then $\mathcal{H}$ is Maker's win.*

The proof is similar ideologically to that of Erdős and Selfridge and uses an appropriately defined potential function. To illlustrate this criterion, consider the arithmetic progression game $W(n, s)$. This is a Maker-Breaker game, whose board is $[n]$, and Maker wins if in the end he claims an arithmetic progression of length s. Let $s(n)$ be the largest s for which Maker wins the game $W(n, s)$. Since the number of arithmetic progressions of length s in n is easily seen to exceed $\frac{n^2}{4(s-1)}$, and each pair of elements $x \neq y \in [n]$ is contained together is at most $\binom{s}{2}$ arithmetic progressions, applying Theorem 4.3 gives $s(n) \geq (1-o(1))\log_2 n$. An application of the Erdős-Selfridge bound for Breaker's side gives $s(n) < 2\log_2 n + 1$. The right asymptotic answer here is actually $s(n) = (1 + o(1))\log_2 n$ [7].

5. Biased games, threshold bias

Many (unbiased) Maker-Breaker games are drastically in favor of Maker, and he wins them easily. Here are few such examples, where the board is the edge set of K_n, we assume n to be large enough. In the triangle game, where Maker wins if in the end he creates a copy of the triangle K_3, Maker is easily seen to win in 4 moves. For the connectivity game, Maker wins as well, say, by Lehman's Theorem 2.7. The *non-planarity game*, where Maker wins if in the end his graph is non-planar, is a complete no-brainer: by Euler's formula, a graph on n vertices with more than $3n - 6$ edges is non-planar, so Maker just waits to accumulate $3n - 5$ edges to declare his victory. If so, it appears quite natural to change the rules of the game to level the field and to increase Breaker's chances to win. One obvious way to do it is to introduce the game bias, as proposed in the pioneering paper of Chvátal and Erdős [19]; this is the subject of this section. Another possible approach is to sparsify the board of the game, and we will discuss it later.

Here is a formal definition of a biased Maker-Breaker game. Let m and b be positive

integers, and let $(X, \mathcal{H})$ be a hypergraph. The *biased* $(m : b)$ *Maker-Breaker game* $(X, \mathcal{H})$ is the same as the Maker-Breaker game $(X, \mathcal{H})$, except that Maker claims m free board elements per move and Breaker claims b elements. The numbers m and b are referred to as the *bias* of Maker and Breaker, respectively. The most frequently considered case is that of $m = 1$.

To illustrate the definition, consider the biased $(1 : b)$ *triangle game* $\mathcal{H}_{K_3,n}$, as was done in [19]. For $b \leq \sqrt{2n} - C$ for some $C > 0$ Maker wins by first accumulating enough edges at a vertex u, and then by closing a triangle containing u. For $b \geq 2\sqrt{n}$, the game is Breaker's win — in response to each move $e_i = (u_i, v_i)$ of Maker, Breaker claims free $b/2$ edges incident to u_i and $b/2$ free edges incident to v_i, also blocking all immediate threats of Maker. The critical value of $b = b(n)$ for this game is still unknown, Balogh and Samotij [6] improved recently Breaker's side to $b \geq (2 - 1/24)\sqrt{n}$.

If we are serious about biased Maker-Breaker games, we should probably start our systematic study from the simplest case where the winning sets of the game hypergraph are pairwise disjoint. This is the famous Box Game introduced by Chvátal and Erdős [19]. In a game $\mathrm{Box}(p, q; a_1, \ldots, a_n)$ the board of the game X is a union of pairwise disjoint sets (boxes) $A_1, \ldots, A_n$ of sizes $a_1, \ldots, a_n$, respectively, forming the game hypergraph. To pay homage to this important game, and also with future games with identity changes in mind, we call players BoxMaker and BoxBreaker. In each move BoxMaker removes p elements from the boxes, and BoxBreaker destroys q boxes of his choice in return. BoxMaker wins if in the end he manages to empty one of the boxes before it is destroyed by BoxBreaker. In the case where all n boxes are of equal size s, we use the notation $\mathrm{Box}(p, q; n \times s)$. This game was analyzed by Chvátal and Erdős for the nearly uniform case, the analysis for the general case was performed by Hamidoune and Las Vergnas [44].

Theorem 5.1 ([19]). *If* $s \leq (p-1)\sum_{i=1}^{n-1} 1/i$*, then BoxMaker, as the first or the second player, wins* $\mathrm{Box}(p, 1; n \times s)$.

Proof. Follows from a bit more general statement we give now. Let $a_1 \leq \ldots a_n \leq a_1 + 1$. Define $f(n, p)$ by the following recursion: $f(1, p) = 0$ and $f(n, p) = \left\lfloor \frac{n(f(n-1,p)+p)}{n-1} \right\rfloor$ for $n \geq 2$. If $\sum_{i=1}^{n} a_i \leq f(n, p)$, then BoxMaker, as the second player, wins $\mathrm{Box}(p,1; a_1,\ldots,a_n)$. The proof proceeds by induction, for the inductive step BoxMaker in his current turn removes p elements to keep the surviving boxes leveled. One can easily show that $f(n, p) \geq (p-1)n\sum_{i=1}^{n-1} 1/i$. □

Theorem 5.2 ([19]). *If* $s > p\sum_{i=1}^{n} 1/i$*, then BoxBreaker wins* $\mathrm{Box}(p, 1; n \times s)$.

Proof. We give a proof from [49]. At any point of the game, denote the set of surviving boxes by S. BoxBreaker always destroys a box $i \in S$ of minimum size. Suppose by contradiction that BoxMaker wins the game at move k, $1 \leq k \leq n$. W.l.o.g. assume that BoxBreaker destroys box i in his ith move, and in his kth move BoxMaker fully claims box k. Let c_i denote the remaining size of box $i \in S \cap \{1, \ldots, k\}$. Define now the potential function Ψ by

$$\Psi(j) := \frac{1}{k-j+1} \sum_{i=j}^{k} c_i \,,$$

the potential just before BoxMaker's move j. Then $\Psi(k) = c_k \leq p$, as BoxMaker wins the game at move k, while $\Psi(1) = s$. In his jth move BoxMaker decreases $\Psi(j)$ by at

most $p/(k-j+1)$; in his jth move BoxBreaker destroys the smallest surviving box. Thus $\Psi(j+1) \geq \Psi(j) - p/(k-j+1)$. It follows that

$$\Psi(k) \geq s - \left(\frac{p}{k} + \frac{p}{k-1} + \cdots + \frac{p}{2}\right) \geq s - p\left(\sum_{i=1}^{n} \frac{1}{i} - 1\right) > p\,,$$

a contradiction. □

We conclude that for the uniform Box Game $\mathrm{Box}(p, q; n \times s)$, the game changes hands around $p = s/\ln n$.

Returning to general biased Maker-Breaker games, let us state a criterion for Breaker's win due to Beck [8], sometimes called the biased Erdős-Selfridge criterion.

Theorem 5.3. *Let X be a finite set, let $\mathcal{H}$ be a family of subsets of X, and let p and q be positive integers. If*

$$\sum_{A \in \mathcal{H}} (1+q)^{-|A|/p} < \frac{1}{1+q}\,, \tag{5.1}$$

then Breaker has a winning strategy in the $(p:q)$ game $(X, \mathcal{H})$.

The proof, while certainly non-trivial, is similar to that of Theorem 4.2. There is an analog of Theorem 4.3 for the biased case [8], but we will not state it here.

Maker-Breaker games are bias monotone. By this we mean the following formal statement: if the $(m:b)$ Maker-Breaker game $(X, \mathcal{H})$ is Maker's win, then so is the $(m:(b-1))$ game. Maker just adapts his winning strategy for the $(m:(b-1))$ game, each time assigning an arbitrary fictitious bth element to Breaker after Breaker's move. This leads us to the following very important definition.

Definition 5.4. Let $(X, \mathcal{H})$ be a hypergraph such that $\min\{|A| : A \in \mathcal{H}\} \geq 2$. The unique positive integer $b_{\mathcal{H}}$ such that Breaker wins the Maker-Breaker $(1:b)$ game $(X, \mathcal{H})$ if and only if $b \geq b_{\mathcal{H}}$ is called the *threshold bias* of $(X, \mathcal{H})$.

Determining or estimating the threshold bias of a game is a central goal of the theory of biased Maker-Breaker games, and is the main subject of this section. For the triangle game $\mathcal{H}_{K_3,n}$, it follows from our prior discussion that the threshold bias $b_{\mathcal{H}}$ is of order $\sqrt{n}$; determining its asymptotic value remains open.

Let us ask ourselves now: for natural biased games on the edge set of K_n, like positive minimum degree, connectivity, Hamiltonicity, etc., what are the values of the threshold bias? How do they compare between themselves? To the reader unexperienced in positional games these questions must appear very challenging, and even making an intelligent guess should be not so easy.

We start with Breaker's side. This is achieved through the following theorem of Chvátal and Erdős [19].

Theorem 5.5. *For every $\epsilon > 0$, all large enough n and $b \geq (1+\epsilon)\frac{n}{\ln n}$, Breaker can isolate a vertex in the $(1:b)$ Maker-Breaker game, played on $E(K_n)$.*

Proof. Breaker first builds a clique C of size $b/2$ such that all vertices of C are isolated in Maker's graph. In his turn i he locates two isolated vertices u_i, v_i in Maker's graph, claims (u_i, v_i) and then joins them completely to the current C, claiming more edges if needed.

Maker in his turn can touch only one vertex of the clique. At the second stage, Breaker's goal is to isolate one of the vertices of C. Observe that all edges inside C have already been taken by Breaker, thus the only relevant edges are those between C and its complement $V \setminus C$; moreover, the edge sets $E_v = \{(u,v) : v \in V \setminus C\}, v \in C$, are disjoint. Thus we can appeal to the box game $\mathrm{Box}(b, 1; \{E_v : v \in C\})$, where Breaker disguises himself as BoxMaker. Applying Theorem 5.1 we derive that Breaker can claim all edges of some E_v, isolating v and winning the game. □

We conclude that the threshold bias for all games on K_n, where all winning sets of Maker are spanning graphs of positive minimum degree, is at most $(1+o(1))n/\ln n$. This might appear as a rather humble beginning, but the truth is that for quite many of them this is a tight estimate!

We now switch to Maker's side. Consider first the connectivity game. Already Chvátal and Erdős showed, probably quite surprisingly, that the threshold bias for this game is of asymptotic order $n/\log n$. Here we present an argument of Beck [8], providing also a better multiplicative constant.

Theorem 5.6. *The threshold bias $b_{\mathcal{C}_n}$ for the connectivity game $\mathcal{C}_n$ on K_n satisfies: $b_{\mathcal{C}_n} \geq (1-o(1))\frac{n}{\log_2 n}$.*

Proof. Let $\epsilon > 0$, and fix $b = b(n) = (1-\epsilon)n/\log_2 n$. We prove that Maker wins $\mathcal{C}_n$ playing against bias b. Observe that in order for Maker to win, it is sufficient (and also necessary) to put an edge into every cut $[S, \bar{S}]$ for $\emptyset \neq S \neq [n]$. So we see another change of roles here: Maker plays as (Cut) Breaker in the *cut game*. The board of the cut game is $E(K_n)$, and the winning sets are exactly the cuts $A_S = [S, \bar{S}]$. Applying criterion (5.1), we need to verify that

$$\sum_S 2^{-|A_S|/b} = \sum_{k=1}^{n/2} \binom{n}{k} 2^{-k(n-k)/b} < \frac{1}{2}\,,$$

which can be done through standard asymptotic manipulations, omitted here. □

What is then the asymptotic value of the threshold bias for the connectivity game and several related games? Which constant should we put in front of $n/\ln n$? Erdős, with an amazing foresight, suggested the following very surprising solution. Suppose both Maker and Breaker in their $(1:b)$ game on $E(K_n)$, instead of being utterly clever and using perfect strategies, play *randomly*. Then the resulting Maker's graph is a random graph on n vertices with m edges for $m = \left\lceil \binom{n}{2}/(b+1) \right\rceil$, i.e., a graph drawn from the probability distribution $G(n,m)$. This puts us in the realm of random graphs, a very developed field where the understanding was far ahead that of positional games. We do not dwell on the background and known results in the theory of random graphs, referring the reader instead to its standard sources [18, 53]. The relevant results are those about the thresholds for positive minimum degree, connectedness, and Hamiltonicity in $G(n,m)$. All three properties are known to appear typically at $m^* = \frac{1}{2}n\ln n$ (much more precise statements are available). This would translate to the threshold bias $b^* \approx \binom{n}{2}/m^* = n/\ln n$ for the random game. Now, the *Erdős paradigm*, or the random graph intuition, suggested that for some biased Maker-Breaker games, like the connectivity game, the threshold bias for the perfectly played games should be asymptotically the same as for the entirely different random games. This approach bridges between two seemingly unrelated fields — positional games and random

graphs, and indicates the very important role of probabilistic considerations in completely deterministic games. A very bold conjecture — which has proven to be true!

Now we state three recent results that established asymptotically the threshold bias for the connectivity game, the minimum degree c game, and the Hamiltonicity game. The first two theorems are due to Gebauer and Szabó [41], the third is due to the author [57].

Theorem 5.7. *For every fixed $\epsilon > 0$ and all sufficiently large n, if $b = (1-\epsilon)\frac{n}{\ln n}$, then Maker wins the $(1:b)$ Maker-Breaker connectivity game $\mathcal{C}_n$.*

Theorem 5.8. *For every fixed $\epsilon > 0$ and every fixed positive integer c, for all large enough n, if $b = (1-\epsilon)\frac{n}{\ln n}$, then Maker can build a spanning graph of minimum degree at least c in the $(1:b)$ Maker-Breaker game played on $E(K_n)$.*

Theorem 5.9. *For every fixed $\epsilon > 0$ and all sufficiently large n, if $b = (1-\epsilon)\frac{n}{\ln n}$, then Maker wins the $(1:b)$ Maker-Breaker Hamiltonicity game played on $E(K_n)$.*

Recalling Theorem 5.5, we conclude that the threshold bias for each of the three games above is asymptotic to $n/\ln n$, completely in line with the Erdős paradigm!

We will not say much about the proofs of the above theorems, referring the reader instead to the original papers. Let us mention that the proofs of the first two theorems go back to basics — Maker reaches his goal directly, instead of using dual approaches and descriptions as in the proof of Theorem 5.6. Cleverly devised potential functions are used in both of the proofs. The third proof uses a modification of the second result and its proof. It turns out that for the minimum degree c game, Maker has a strategy to reach degree c at any vertex v before Breaker accumulates $(1-\delta)n$ edges at v, for some $\delta = \delta(\epsilon) > 0$. The strategy of Maker, as given by [41], points Maker to a vertex v (specified by the current situation on the board) and tells him to claim an *arbitrary* free edge incident to v. The crucial twist is to use *randomness* here and to choose instead a *random* free edge at v, out of at least δn edges available. One can argue that following this random strategy, with positive probability Maker can create a pretty strong expander in linearly many moves. At this point the deterministic nature of the game comes to our help — the game considered is of perfect information, and thus winning with positive probability against a given strategy of Breaker means there is a deterministic (but unspecified) strategy to win. Returning to the Hamiltonicity game, Maker then quickly turns his expander into a connected graph and finally augments his connected expander to a Hamiltonian graph in a linear number of moves; here the proof uses fairly standard techniques from the theory of random graphs. Altogether, Maker wins the game in at most $18n$ moves, when the board is still mostly empty.

Now we cite another important result about biased Maker-Breaker games, due to Bednarska and Łuczak [15]. For a given graph H, Maker wins the *H-game* played on the edges of a host graph G if in the end he possesses a copy of H. For the case $H = K_3$ and $G = K_n$ we get the above treated triangle game. Define now the maximum 2-density $m_2(H)$ of H, a frequently used notation in random graphs, by

$$m_2(H) = \max_{H_0 \subseteq H, |V(H_0)|>2} \frac{|E(H_0)|-1}{|V(H_0)|-2}\,.$$

Theorem 5.10. *Let H be a graph with at least three non-isolated vertices. The threshold bias for the Maker-Breaker H-game on $E(K_n)$ satisfies: $b = \Theta\left(n^{1/m_2(H)}\right)$.*

The proof of Maker's side uses a random strategy again, this time in the simplest possible form: Maker chooses each time a random edge to claim. This shows yet again that probabilistic considerations and arguments are ubiquitous in positional games, quite a surprising phenomenon.

Yet another connection between positional games and randomness was revealed in [34], where Maker's win in certain biased games was achieved through Maker creating a random graph with few edges deleted at each vertex, and then invoking results about local resilience of random graphs [69].

We conclude this section with a very entertaining argument of Beck [10], providing a lower bound for the Ramsey number $R(3,t)$ through biased Maker-Breaker games. The bound obtained is superseded by the best possible bound $R(3,t) = \Omega(t^2/\log t)$ of Kim [54] (see recent [17, 35] for better constants), but it matches the best bound known for long 35 years, obtained through various approaches [25, 28, 56, 68].

Theorem 5.11. *There exists a constant $c > 0$ such that $R(3,t) \geq ct^2/\ln^2 t$.*

Proof. Set $b = 2\sqrt{n}$. Imagine two players playing on the edges of K_n. The first player, taking b edges at a time, thinks of himself as Breaker in the $(1:b)$ Maker-Breaker game, whose goal is to prevent Maker from claiming a triangle. The second player, claiming one edge per move, thinks of himself as Breaker in a different game, namely, the $(b:1)$ Maker-Breaker game, whose goal is to claim an edge in every vertex subset of cardinality $t = C\sqrt{n}\ln n$. The first player wins his game by our analysis of the triangle game. The second player is victorious too for large enough C — this can be derived by applying the biased Erdős-Selfridge criterion (5.1) (the calculations are omitted). The result of the game, or perhaps of the games, is hence a partition of $E(K_n)$ into two graphs, where the first graph is triangle-free, and the second graph does not have a clique of size t. It thus follows that $R(3,t) > n$. □

6. Avoider-Enforcer games

Recall the game of Sim described in Section 2. It has the interesting feature — the player who occupies a winning set first actually loses. This is an example of *reverse*, or misére-type, games. Games of this type are the subject of this section. Reverse games are certainly interesting for their own sake, but for those who seek additional motivation to consider them, we now give an example of a Maker-Breaker game, where Maker relies on a reverse game to ensure his win.

Theorem 6.1 ([45]). *For every fixed $\epsilon > 0$ and all sufficiently large n, if $b = (\frac{1}{2} - \epsilon)n$, then Maker wins the $(1:b)$ Maker-Breaker non-planarity game on $E(K_n)$.*

Proof. It follows from Euler's formula that for $k \geq 3$, if a graph G on n vertices has more than $\frac{k}{k-2}(n-2)$ edges and no cycles of length shorter than k, then G is non-planar. Let $\alpha = \frac{\epsilon}{1-2\epsilon}$, and let k be the smallest integer satisfying $1 + \alpha > \frac{k}{k-2}$. In order the ensure his final graph is non-planar, it suffices for Maker to *avoid* creating cycles shorter than k in his first $(1+\alpha)n$ moves. This is an easy task — the board will still have $\Theta(n^2)$ free edges after this number of moves, due to our choice of α. Maker always chooses his next edge so as not to close a cycle of length less than k and not to create a vertex of degree at least $n^{1/k}$; showing this is a feasible strategy is an easy exercise. □

We now define Avoider-Enforcer games formally. Let a and b be positive integers, and let $(X, \mathcal{H})$ be a hypergraph. In the *biased* $(a : b)$ *Avoider-Enforcer game* $(X, \mathcal{H})$ the two players are Avoider and Enforcer, with Avoider moving first and claiming exactly a free elements in each turn, while Enforcer claims exactly b free elements in each turn. Enforcer wins the game if he forces Avoider into claiming fully one of the sets $A \in \mathcal{H}$, and Avoider wins otherwise. The members A of the game hypergraph $\mathcal{H}$ are sometimes called *losing sets*, to reflect the nature of the game.

Having seen the important role of the bias monotonicity and the threshold bias in Maker-Breaker games, we can hope that Avoider-Enforcer games behave similarly. However, this is pretty much *not* the case, as has been observed in [50]. Consider the following simple example: the game hypergraph $\mathcal{H}$ consists of two disjoint sets of size 2 each. It is immediate to check that for $a = b = 1$ Avoider is the winner, for $a = 1, b = 2$ Enforcer wins, and for $a = b = 2$ Avoider wins again; the example can easily be generalized to larger sets and bias numbers. This *lack of monotonicity* is a fairly disturbing feature, which prompted the authors of [47] to adjust the definition in the following, rather natural, way: now Avoider claims *at least* a elements in each turn, while Enforcer claims *at least* b elements. This version is easily seen to be bias monotone, and for this reason we call this set of rules *monotone rules*, while the former set of rules is called *strict rules*. Each monotone Avoider-Enforcer game $\mathcal{H}$ has the threshold bias $b^{\text{mon}}_{\mathcal{H}}$, which is the largest non-negative integer b for which Enforcer wins the corresponding $(1 : b)$ game. For strict rules, we can define instead the *lower threshold bias* $b^{-}_{\mathcal{H}}$ as the largest integer such that Enforcer wins the $(1 : b)$ game for every $b \le b^{-}_{\mathcal{H}}$, and the *upper threshold bias* $b^{+}_{\mathcal{H}}$ as smallest integer such that Avoider wins the $(1 : b)$ game for every $b > b^{+}_{\mathcal{H}}$.

Just like for Maker-Breaker games, determining or estimating the threshold bias(es) is a central task for Avoider-Enforcer games, under both sets of rules. The difference is that here the situation is frequently much more challenging, and our understanding of Avoider-Enforcer games does not quite match that of their Maker-Breaker counterparts.

As a warm-up example, consider the game $\mathcal{H}_{P_2,n}$, where Avoider aims to avoid creating a copy of the path P_2 of two edges. The biases for this game are as follows [47]: $b^{+}_{\mathcal{H}_{P_2,n}} = \binom{n}{2} - 2$, $b^{-}_{\mathcal{H}_{P_2,n}} = \Theta\left(n^{3/2}\right)$ and $b^{\text{mon}}_{\mathcal{H}_{P_2,n}} = \binom{n}{2} - \left\lfloor \frac{n}{2} \right\rfloor - 1$.

One may be tempted to think that for Avoider-Enforcer games on a game hypergraph $\mathcal{H}$, the monotone threshold bias $b^{\text{mon}}_{\mathcal{H}}$ is always between the lower and the upper threshold biases $b^{-}_{\mathcal{H}}$ and $b^{+}_{\mathcal{H}}$. This is not the case as we will see shortly.

We now state several results obtained for both game rules. Let us start with the strict rules.

Theorem 6.2 ([50]). *For the Avoider-Enforcer connectivity game $\mathcal{C}_n$ on $E(K_n)$, played under strict rules, we have $b^{-}_{\mathcal{C}_n} = b^{+}_{\mathcal{C}_n} = \left\lfloor \frac{n-1}{2} \right\rfloor$.*

This is a very unusual result in the sense that is provides the *exact* value of the threshold biases. Avoider's side $b^{+}_{\mathcal{C}_n}$ is trivial — if Avoider ends up with less than $n-1$ edges, he just cannot create a spanning tree and is thus guaranteed to win. Enforcer's side uses the well-known fact that K_n contains $\lfloor n/2 \rfloor$ edge-disjoint spanning trees and the following theorem.

Theorem 6.3 ([50]). *If G contains $b+1$ pairwise edge-disjoint spanning trees, then Enforcer wins the $(1 : b)$ Avoider-Enforcer connectivity game played on $E(G)$ under strict rules.*

This theorem can be considered to be the Avoider-Enforcer analog of Lehman's Maker-Breaker Theorem 2.7. Peculiarly enough, the biased Maker-Breaker version of Lehman does

not go through: for any k one can construct an example of a graph G with k edge-disjoint spanning trees, where Breaker wins the $(1:2)$-connectivity game. Theorem 6.3 also readily implies the following result.

Theorem 6.4 ([50]). *For the k-connectivity Avoider-Enforcer game $\mathcal{C}_n^k$ on K_n, $k \geq 2$, one has: $\frac{n}{2k} \leq b^-_{\mathcal{C}_n^k} \leq b^+_{\mathcal{C}_n^k} \leq \frac{n}{k}$.*

For the Hamiltonicity game we understand Enforcer's side well under both rules:

Theorem 6.5 ([59]). *If $b \leq (1-o(1))\frac{n}{\ln n}$, then Enforcer has a winning strategy in the $(1:b)$ Hamiltonicity game on $E(K_n)$, under either strict or monotone rules.*

Moving on to the monotone rules, we have the following key result.

Theorem 6.6 ([47]). *If $b \geq (1+o(1))\frac{n}{\ln n}$, then Avoider has a strategy to be left with an isolated vertex in the monotone $(1:b)$ game on $E(K_n)$.*

Theorems 6.5 and 6.6 combined establish the threshold bias b^{mon} of several important games (connectivity, perfect matching, Hamiltonicity, etc.) to be asymptotically equal to $n/\ln n$. Observe that for the strict connectivity game both threshold biases, which happen to be equal by Theorem 6.2, are quite far from the threshold bias for the monotone version, perhaps contrary to our intuition.

Finally, Clemens et al. [20] showed very recently:

Theorem 6.7. *For n large enough and $b \geq 200n\ln n$, Avoider has a strategy to be left with a graph with at most one cycle in the $(1:b)$ game on $E(K_n)$, under either strict or monotone rules.*

It follows that the threshold biases under monotone rules, and the upper threshold biases $b^+_{\mathcal{H}}$ under strict rules for the non-planarity and the non-k-colorability games with $k \geq 3$ are at most $200n\ln n$. Still, for these games we are quite far from nailing these biases.

For Avoider-Enforcer games with losing sets of constant size the situation is very far from satisfactory — we know few sporadic results or bounds. For example, for the triangle game $\mathcal{H}_{K_3,n}$ the monotone threshold bias is $\Theta(n^{3/2})$ [47], very far from the threshold bias for the Maker-Breaker version; for the strict game we have: $b^-_{\mathcal{H}_{K_3,n}} = \Omega(n^{1/2})$, $b^+_{\mathcal{H}_{K_3,n}} = O(n^{3/2})$ [14]. A recent paper of Bednarska-Bzdęga [14] contains several interesting results of this sort.

As for general tools to tackle Avoider-Enforcer games, we are somewhat shorthanded here. We have the following results:

Theorem 6.8 ([50]). *If $\sum_{A\in\mathcal{H}}(1+1/a)^{-|A|+a} < 1$, then Avoider wins the biased $(a:b)$ game $\mathcal{H}$, under both strict and monotone rules, for every $b \geq 1$.*

(Observe the lack of sensitivity to the value of b in the above criterion — an obvious drawback.)

Theorem 6.9 ([14]). *Let $(X,\mathcal{H})$ be a hypergraph with all sets $A \in \mathcal{H}$ of size at most r. If $\sum_{A\in\mathcal{H}}(1+b/(ar))^{-|A|+a} < 1$, then Avoider wins the biased $(a:b)$ game $\mathcal{H}$, under both strict and monotone rules.*

This criterion is handy for games with losing sets of small size.

7. More boards, more games

Most of the concrete games we have considered so far are played on the complete graph K_n. This does not have to be the case, and many games on sparser graphs are equally interesting. Also, sparsifying the game board (rather than turning to biased games) can be seen as an alternative approach to provide Breaker with higher chances to win standard graph games against Maker.

A very typical setting is games on random graphs. In this scenario, for a given game type, say Maker-Breaker Hamiltonicity, we first set up a probability space of graphs, say, the binomial random graphs $G(n,p)$, and then ask about the probability of generating a board which is a win of each of the players.

Here is a representative result of this sort. Denote $N = \binom{n}{2}$. Consider the random graph process $\tilde{G} = (G_i)_{i=0}^{N}$ on n vertices, described as follows. Start with the empty graph G_0 with vertex set $[n]$, and then for $1 \le i \le N$, form a graph G_i by adding to G_{i-1} a random missing edge. For a monotone graph property P and a (random) graph process $\tilde{G}$, we define the *hitting time* $\tau(\tilde{G}, P)$ as the minimal i such that G_i, the ith graph of the process, possesses P. For a random graph process $\tilde{G}$, the hitting time $\tau(\tilde{G}, P)$ becomes a random variable and we can study its typical behavior. Consider now the unbiased Hamiltonicity game $\mathcal{HAM}$ with Breaker moving first, and let $M_{\mathcal{HAM}}$ be the property "Maker wins $\mathcal{HAM}$ on G". Breaker clearly wins for the empty graph G_0, and by the classical Chvátal-Erdős result [19], Maker is the winner for $G_N = K_n$. So the hitting time $\tau(\tilde{G}, P)$ lies somewhere in between. Those familiar with the theory of random graphs can guess that the key to Maker's win will be the disappearance of vertices of small degree. Indeed, if $\delta(G) < 4$, then Breaker, moving first, wins the game by claiming all but at most one edge at a vertex of minimum degree in G. Thus, $\tau(\tilde{G}, M_{\mathcal{HAM}}) \ge \tau(\tilde{G}, \delta_4)$, where δ_4 is the property of having minimum degree at least 4.

Theorem 7.1 ([16]). *For a random graph process $\tilde{G}$ on n vertices, with high probability $\tau(\tilde{G}, M_{\mathcal{HAM}}) = \tau(\tilde{G}, \delta_4)$.*

Thus, typically in the random graph process, Maker starts winning the Hamiltonicity game *exactly* at the moment the last vertex of degree less than 4 disappears. From this result one can derive, in a rather standard way, the corresponding result for $G(n,p)$ — the threshold probability for Maker's win stands at $p(n) = \frac{\ln n + 3\ln\ln n}{n}$. The same paper [16] gets hitting time results also for the perfect matching and the k-connectivity Maker-Breaker games. Biased Hamiltonicity games on $G(n,p)$ were considered in [31], where it was shown that for $p \gg \frac{\log n}{n}$, the threshold bias $b_{\mathcal{HAM}}$ satisfies typically $b_{\mathcal{HAM}} = (1 + o(1))\frac{np}{\log n}$. This fits very well with the probabilistic intuition of Erdős, as was predicted by Stojaković and Szabó [70] who initiated the study of positional games on random graphs.

The H-game, where Maker wins if he creates a copy of a fixed graph H, was considered in its unbiased version for the case of random boards in [70], [62], [63]. In particular, Müller and Stojaković proved in [62]:

Theorem 7.2. *Let $k \ge 4$ be fixed. There exists a constant $c = c(k) > 0$ such that for $p \le cn^{-\frac{2}{k+1}}$, a random graph $G \sim G(n,p)$ is with high probability such that Breaker wins the unbiased Maker-Breaker K_k-game on G.*

The matching result for starting Maker (if $p \ge Cn^{-\frac{2}{k+1}}$, then Maker typically wins the K_k-game on $G(n,p)$) can be derived from the general Ramsey-type result of Rödl and

Ruciński [65] and the strategy stealing argument, and it can be easily adapted for the case where Breaker starts. The case $k = 3$ turns out to be different. There the threshold lies at $p = n^{-5/9}$ for a good local reason: Maker wins the triangle game on the graph $K_5 - e$, hence Maker wins no later than this graph appears in $G(n,p)$ – which typically happens at $p = n^{-5/9}$. The paper [62] provides a hitting time version of this result.

Another sparsification-type approach asks for the minimal size of a game board, on which Maker can create a prescribed structure. For example, [48] and [5] proved that a graph G on n vertices on which Maker wins the positive minimum degree game has at least about $10n/7$ edges, and this estimate is tight. Gebauer, building partly on ideas from [29], showed:

Theorem 7.3 ([40]). *For every $d > 0$ there exists $c = c(d)$ such that for every graph H on n vertices of maximum degree at most d, there is a graph G with at most cn edges such that Maker wins the unbiased H-game on G.*

These problems can be viewed as game versions of size Ramsey numbers. For a graph H, the *size Ramsey number* $\hat{r}(H)$ is the smallest M for which there exists a graph G with M edges such that any red-blue coloring of the edges of G produces a monochromatic copy of H (we say also that G arrows H). If G arrows H, then by the strategy stealing argument (again!), Maker as the first player wins the H-game on G. Hence size Ramsey numbers upper bound their game counterparts. The so obtained bounds are not always tight: for example, Rödl and Szemerédi showed [66] the existence of a graph H on n vertices of maximum degree 3 and size Ramsey number $\hat{r}(H) \geq cn \log^{\frac{1}{60}} n$, thus making it impossible to derive Theorem 7.3 from general size Ramsey results.

In many games, the identity of the winner is easy to establish, and one can then ask how long it takes him to win. For example, Lehman's Theorem 2.7 and its proof show that if a graph G on n vertices has two edge-disjoint spanning trees, then Maker, as the first or the second player, wins the connectivity game on G in $n-1$ moves; this is clearly optimal. Define the *move number* $\text{move}(\mathcal{H})$ of a weak (resp. strong) game $\mathcal{H}$ as the smallest t such that Maker (respectively FP) has a strategy to win $\mathcal{H}$ within t moves. If Breaker is the winner (SP draws, resp.), we set $\text{move}(\mathcal{H}) = \infty$. The move number for Avoider-Enforcer games is defined accordingly. Here is a sample of results about this concept. [46] showed that for sufficiently large n, playing on the edges of K_n, Maker can create a Hamilton path in $n-1$ moves, and a perfect matching (assuming n is even) in $n/2+1$ moves. Hefetz and Stich proved:

Theorem 7.4 ([52]). *Let $\mathcal{H}AM$ denote the Maker-Breaker Hamiltonicity game played on the edges of K_n. Then* $\text{move}(\mathcal{H}AM) = n+1$*, for all sufficiently large n.*

For the Maker-Breaker clique game (K_n, K_q), the move number is known to be between $2^{\frac{q}{2}}$ [10] and $2^{\frac{2q}{3}}\text{poly}(q)$ [39], for all large enough n. For Avoider-Enforcer games, Anuradha et al. [4] proved that Avoider can stay planar for as long as $3n-26$ turns when playing against Enforcer in the unbiased game on K_n, a constant away from the trivial upper bound of $3n-6$.

Fast wins are closely related to strong games. Observe that if in a game $\mathcal{H}$ all winning sets have cardinality n, and in the weak game on $\mathcal{H}$ Maker has a strategy to win in n moves, then in the strong game on $\mathcal{H}$ First Player wins in n moves. Thus for example we can derive that FP wins the Hamilton path game on K_n. This scenario is quite rare though. Still, in a

recent exciting development, Ferber and Hefetz were able to use fast winning strategies in Maker-Breaker games to analyze much harder strong games.

Theorem 7.5 ([32, 33]). *Let $\mathcal{PM}$, $\mathcal{HAM}$, $\mathcal{C}^k$ denote the perfect matching, Hamiltonicity and spanning k-connectivity games, resp., played on the edges of K_n. The strong versions of these games are First Player's win, for large enough n. Also,* $\mathrm{move}(\mathcal{PM}) \leq \frac{n}{2} + 2$, $\mathrm{move}(\mathcal{HAM}) \leq n + 2$, *and* $\mathrm{move}(\mathcal{C}^k) = \left\lfloor \frac{kn}{2} \right\rfloor + 1$ *for* $k \geq 3$.

The key to all these proofs is analysis of fast strategies for the corresponding Maker-Breaker games and their adaptation for strong games. A natural outcome of this proof approach is explicit winning strategies for FP. These results provide a new lease of life for the whole subject of strong games, notorious for its difficulty.

Discrepancy games can be viewed as a hybrid between Maker-Breaker and Avoider-Enforcer games. In such games, the first player, called Balancer, aims to end up with the correct proportion of elements in each winning set. For example, in the unbiased case Balancer strives to get about half of his elements (not much more, not much less) in every winning set. These games have been considered in [2, 51], see also Chapters 16, 17 of [11]. [36] describes a strategy for Balancer to construct a pseudo-random graph; since pseudo-random graphs are known to have many nice features [58], this result guarantees Maker's win in a variety of games.

There are also games involving directed/oriented graphs, or graph orientation. For example, in the tournament game a tournament T on k vertices is given, and Maker and Breaker take turns in claiming free edges of K_n, one edge each, with Maker also orienting his edges. Maker wins the game if his graph contains in the end a copy of T. Clemens, Gebauer, and Liebenau proved [21] that for $k = (2 - o(1)) \log_2 n$ Maker can create a copy of any given k-vertex tournament T. This is asymptotically optimal and resolves a problem posed by Beck in [11].

Finally, let us describe yet another quite interesting class of games, not very well studied as of yet. In these games, the players are Picker and Chooser. Picker in his ith turn picks two free elements v_i, v_i' of the board X and presents them to Chooser — who chooses one of these elements, with the other one going to Picker. In the *Chooser-Picker* version of the game, Chooser wins if in the end he completes a winning set $A \in \mathcal{H}$, and Picker wins otherwise; in the *Picker-Chooser* version Picker wins if he occupies a winning set. These games, especially the Chooser-Picker variant, appear to be similar to Maker-Breaker versions, and it was even conjectured that if Breaker wins the game on $\mathcal{H}$, then Picker, whose job appears to be even easier, wins the Chooser-Picker game on $\mathcal{H}$; however, this was disproved in [55]. Still, Bednarska-Bzdęga showed [13] that the Erdős-Selfridge criterion (4.1) is also a winning criterion for Picker in Chooser-Picker games. Beck in [11] analyzed the Picker-Chooser clique game (K_n, K_q) with an amazing degree of precision, proving in particular that Chooser starts winning at $q = (1 - o(1))2 \log_2 n$. One can also consider the biased versions of both game types, where at each round Picker picks between $1 + p$ and $p + q$ free elements of the board, Chooser keeps p of them, and the rest goes to Picker. See [11, 13, 22, 23] for more results.

8. Open problems and challenges

In this section we present a representative sample of open problems in the field, and also indicate promising research directions.

For strong games, there are still many more problems than answers, though the situation is more hopeful now after the recent results stated in Theorem 7.5. One concrete example is the following question:

Problem 8.1. *Show that for every positive q there exist t and n_0 such that for every $n \geq n_0$ First Player can win in at most t moves the strong clique game (K_n, K_q).*

In other words, we are essentially asking for an *explicit* winning strategy for FP in the clique game. As we stated before, for $n \geq R(q,q)$ First Player wins due to strategy stealing, but this is highly inexplicit. The problem appears to be open even for the case $q = 5$.

For weak (Maker-Breaker) games, we are in a much better shape. Still there are many nice problems to tackle. In the *degree game*, a graph G of minimum degree d is given, and Maker aims to create a graph of highest possible minimum degree. Since the edges will be split evenly between the players in the end, the best Maker can hope for is $d/2$. Getting around $d/4$ is fairly easy, here is a sketch. Assume for simplicity all degrees in G are divisible by 4. Using an Eulerian orientation, one can split the edges of G between its vertices so that a vertex v gets a set E_v of $d(v)/2$ incident edges assigned to it. Maker then plays a pairing game on the disjoint sets E_v, answering Breaker's move $e \in E_v$ by $e' \in E_v$, and thus guaranteeing a quarter degree at every vertex. At present, $d/4$ is the best known result, even improving it to $(1/4 + \epsilon)d$ would be quite nice. If we allow d to depend on the number of vertices $n = |V(G)|$, then for $d \gg \log n$ we can get $(1/2 - o(1))d$ in the degree game using known discrepancy results, say [2].

Here is a very cute problem due to Duffus, Łuczak, and Rödl [24].

Problem 8.2. *Prove that for integers $b \geq 1$ and $r \geq 3$, there exists $C = C(b,r)$ such that for every graph $G = (V,E)$ of chromatic number at least C, Maker has a strategy to create a graph M of chromatic number at least r when playing the $(1:b)$ biased Maker-Breaker game on $E(G)$.*

(Duffus et al. [24] asked actually the equally interesting vertex version of this problem, where players claim vertices of G rather than edges; we state it in the edge version, more in line with the prevailing setting of Maker-Breaker games.) The unbiased case $b = 1$ is easy – one can take $C = (r-1)^2 + 1$, and the argument goes as follows. Recall first that for every edge decomposition $E(G) = E(M) \cup E(B)$ one has $\chi(G) \leq \chi(M)\chi(B)$, where $\chi(G)$ denotes the chromatic number of G. Hence, if for a graph G with $\chi(G) > (r-1)^2$ Breaker has a strategy to prevent Maker from reaching $\chi(M) \geq r$, then that very strategy guarantees Breaker a graph B with $\chi(B) \geq r$. Maker, who is assumed to move first, can then steal this strategy and achieve $\chi(M) \geq r$. However, even the next case $b = 2, r = 3$ is open. Only very partial results are available, see [1, 30].

The *neighborhood conjecture* is one of the most important problems in positional games, with several ramifications in other combinatorial and computer science questions. Recall that the Erdős-Selfridge criterion (4.1) guarantees that a k-uniform hypergraph $\mathcal{H}$ with less that 2^{k-1} edges is Breaker's win, and is thus 2-colorable by Proposition 4.1. As we mentioned, this is tight for games. The obvious drawback of this result is that it does not take into account the local structure of $\mathcal{H}$, hence taking any non-empty Breaker's win $\mathcal{H}$ and repeating it enough times will eventually create a hypergraph violating the Erdős-Selfridge condition (still Breaker's win of course). A local version of Breaker's winning criterion would be:

Problem 8.3. *Determine*

$$D(k) := \min\{d : \exists\ k\text{-uniform Maker's win } \mathcal{H} \text{ with } \Delta(\mathcal{H}) \leq d\}\,,$$

where $\Delta(\mathcal{H})$ is the maximum degree of $\mathcal{H}$. The definition of $D(k)$ implies in particular that any k-uniform $\mathcal{H}$ with $\Delta(\mathcal{H}) < D(k)$ is 2-colorable. The relation to the famous Lovász Local Lemma [26] is apparent — a standard application of the Local Lemma gives that if $\Delta(\mathcal{H}) \leq \frac{2^{k-1}}{ek}$, then $\mathcal{H}$ is 2-colorable. The gap between known lower and upper bounds for $D(k)$ is truly astonishing. For the lower bound, we know the trivial $D(k) \geq \frac{k}{2} + 1$ (pairing strategy). Gebauer proved [38] that $D(k) \leq (1 + o(1))\frac{2^k}{k}$, thus disproving the original neighborhood conjecture, stated in Beck's book [11]. This bound was improved further to $D(k) \leq (1 + o(1))\frac{2^k}{ek}$ by Gebauer, Szabó and Tardos [42]. The same paper [42] reveals exciting connections between the neighborhood conjecture, the Local Lemma, and the satisfiabilty problem, most central in computer science.

Another problem about Maker-Breaker games due to Beck [11] is as follows.

Problem 8.4. *For the $(m : m)$ Maker-Breaker clique game on K_n, what is the largest clique size q Maker is guaranteed to achieve?*

For the unbiased case $m = 1$ the answer is $q = (1 - o(1))2\log_2 n$, due to Beck [11]. This matches the probabilistic intuition very well, as the clique number of the random graph $G(n, 1/2)$ is typically equal to $(1 - o(1))2\log_2 n$. If so, one can expect that for $m > 1$ the answer should be similar. This is however not the case for $m \geq 6$, as shown in [39].

For Avoider-Enforcer games, the current state of affairs does not quite match the situation for their Maker-Breaker analogs. In particular, we do not know yet to resolve the strict Hamiltonicity game:

Problem 8.5. *What are the threshold biases b^- and b^+ for the $(1 : b)$ Avoider-Enforcer Hamiltonicity game on $E(K_n)$, played under strict rules?*

The gap between the lower bound $b^- \geq (1 - o(1))n/\ln n$ from Theorem 6.5 and the trivial upper bound $b^+ \leq n/2 - 1$ is quite annoying. It would be also very nice to develop a general theory of Avoider-Enforcer H-games, for fixed H; so far these games have mostly been attacked on a game-to-game basis.

Finally, we mention Chooser-Picker and Picker-Chooser games. They are largely an uncharted territory, and natural and attractive problems abound there.

Acknowledgements. The author wishes to thank Asaf Ferber, Dan Hefetz, Miloš Stojaković and Tibor Szabó for careful reading of the manuscript and many helpful comments, and also for an extensive and fruitful cooperation in research on this fascinating subject.

References

[1] N. Alon, D. Hefetz, and M. Krivelevich, *Playing to retain the advantage*, Combin. Probab. Comput. **19** (2010), 481–491.

[2] N. Alon, M. Krivelevich, J. Spencer, and T. Szabó, *Discrepancy games*, Electron. J. Combin. **12** (1) (2005), publ. R51.

[3] N. Alon and J. Spencer, *The Probabilistic Method*, 3rd ed., Wiley, 2008.

[4] V. Anuradha, C. Jain, J. Snoeyink, and T. Szabó, *How long can a graph be kept planar?*, Electron. J. Combin. **15** (1) (2008), publ. N14.

[5] J. Balogh and A. Pluhár, *The positive minimum degree game on sparse graphs*, Electron. J. Combin. **19** (2012), Publ. 22.

[6] J. Balogh and W. Samotij, *On the Chvátal-Erdős triangle game*, Electron. J. Combin. **18** (2011), publ. 72.

[7] J. Beck, *van der Waerden and Ramsey type games*, Combinatorica **1** (1981), 103–116.

[8] _____, *Remarks on positional games*, Acta Math. Acad. Sci. Hungar. **40** (1982), 65–71.

[9] _____, *Deterministic graph games and a probabilistic intuition*, Combin. Probab. Comput. **3** (1994), 13–26.

[10] _____, *Ramsey games*, Discrete Math. **249** (2002), 3–30.

[11] _____, *Combinatorial Games: Tic-Tac-Toe Theory*, Encyclopedia of Mathematics and Its Applications 114, Cambridge University Press, 2008.

[12] _____, *Inevitable randomness in discrete mathematics*, University Lecture Series, 49. Amer. Math. Soc., Providence, RI, 2009.

[13] M. Bednarska-Bzdęga, *On weight function methods in Chooser-Picker games*, Theor. Comput. Sci. **475** (2013), 21–33.

[14] _____, *Avoider-Forcer games on hypergraphs with small rank*, Electron. J. Combin. **21** (1) (2014), publ. P1.2.

[15] M. Bednarska and T. Łuczak, *Biased positional games for which random strategies are nearly optimal*, Combinatorica **20** (2000), 477–488.

[16] S. Ben-Shimon, A. Ferber, D. Hefetz, and M. Krivelevich, *Hitting time results for Maker-Breaker games*, Random Struct. Alg. **41** (2012), 23–46.

[17] T. Bohman and P. Keevash, *Dynamic concentration of the triangle-free process*, submitted, arXiv1302.5963 [math.CO].

[18] B. Bollobás, *Random Graphs*, Academic Press, London, 1985.

[19] V. Chvátal and P. Erdős, *Biased positional games*, Annals Discrete Math. **2** (1978), 221–228.

[20] D. Clemens, J. Ehrenmüller, Y. Person, and T. Tran, *Keeping Avoider's graph almost acyclic*, submitted, arXiv1403.1482 [math.CO].

[21] D. Clemens, H. Gebauer, and A. Liebenau, *The random graph intuition for the tournament game*, submitted, arXiv1307.4229 [math.CO].

[22] A. Csernenszky, *The Picker-Chooser diameter game*, Theor. Comput. Sci. **411** (2010), 3757–3762.

[23] A. Csernenszky, C. I. Mándity, and A. Pluhár, *On Chooser-Picker positional games*, Discrete Math. **309** (2009), 5141–5146.

[24] D. Duffus, T Łuczak, and V. Rödl, *Biased positional games on hypergraphs*. Studia Sci. Math. Hungar. **34** (1998), 141–149.

[25] P. Erdős, *Graph theory and probability II*, Can. J. Math., **13** (1961), 346–352.

[26] P. Erdős and L. Lovász, *Problems and results on 3-chromatic hypergraphs and some related questions*, In: Infinite and finite sets II, Colloq. Math. Soc. J. Bolyai, Vol. 10, North-Holland (1975), 609–627.

[27] P. Erdős and J. L. Selfridge, *On a combinatorial game*, J. Combin. Th. Ser. A **14** (1973), 298–301.

[28] P. Erdős, S. Suen, and P. Winkler, *On the size of a random maximal graph*, Random Struct. Alg. **6** (1995), 309–318.

[29] O. Feldheim and M. Krivelevich, *Winning fast in sparse graph construction games*, Combin. Probab. Comput. **17** (2008), 781–791.

[30] A. Ferber, R. Glebov, M. Krivelevich, H. Liu, C. Palmer, T. Valla, and M. Vizer, *The biased odd cycle game*, Electron. J. Combin. **20** (3) (2013), publ. P9.

[31] A. Ferber, R. Glebov, M. Krivelevich, and A. Naor, *Biased games on random boards*, Random Struct. Alg., to appear.

[32] A. Ferber and D. Hefetz, *Winning strong games through fast strategies for weak games*, Electron. J. Combin. **18** (1) (2011), publ. 144.

[33] _____, *Weak and strong k-connectivity games*, Europ. J. Combin. **35** (2014), 169–183.

[34] A. Ferber, M. Krivelevich, and H. Naves, *Generating random graphs in biased Maker-Breaker games*, submitted, arXiv1310.4096 [math.CO].

[35] G. Fiz Pontiveros, S. Griffiths, and R. Morris, *The triangle-free process and $R(3,k)$*, submitted, arXiv1302.6279 [math.CO].

[36] A. Frieze, M. Krivelevich, O. Pikhurko, and T. Szabó, *The game of JumbleG*, Combin. Probab. Comput. **14** (2005), 783–793.

[37] D. Gale, *The game of Hex and the Brouwer fixed-point theorem*, Amer. Math. Monthly **86** (1979), 818–827.

[38] H. Gebauer, *Disproof of the Neighborhood Conjecture with Implications to SAT*, Combinatorica **32** (2012), 573–587.

[39] _____, *On the Clique-Game*, Europ. J. Combin. **33** (2012), 8–19.

[40] _____, *Size Ramsey number of bounded degree graphs for games*, Combin. Probab. Comput. **22** (2013), 499–516.

[41] H. Gebauer and T. Szabó, *Asymptotic random graph intuition for the biased connectivity game*, Random Struct. Alg. **35** (2009), 431–443.

[42] H. Gebauer, T. Szabó, and G. Tardos, *The Local Lemma is tight for SAT*, 22nd Annual Symposium on Discrete Algorithms (SODA 2011), 664–674.

[43] A. W. Hales and R. I. Jewett, *Regularity and positional games*, Trans. Amer. Math. Soc. **106** (1963), 222–229.

[44] Y. O. Hamidoune and M. Las Vergnas, *A solution to the box game*, Discrete Math. **65** (1987), 157–171.

[45] D. Hefetz, M. Krivelevich, M. Stojaković, and T. Szabó, *Planarity, colorability and minor games*, SIAM J. Discrete Math. **22** (2008), 194–212.

[46] ———, *Fast winning strategies in Maker-Breaker games*, J. Combin. Th. Ser. B **99** (2009), 39–47.

[47] ———, *Avoider – Enforcer: the rules of the game*, J. Combin. Th. Ser. A **117** (2010), 152–163.

[48] ———, *Global Maker-Breaker games on sparse graphs*, Europ. J. Combin. **32** (2011), 162–177.

[49] ———, *Positional Games* (Oberwolfach Seminars), Birkhäuser, 2014.

[50] D. Hefetz, M. Krivelevich, and T. Szabó, *Avoider-Enforcer games*, J. Combin. Th. Ser. A **114** (2007), 840–853.

[51] ———, *Bart-Moe games, JumbleG and discrepancy*, Europ. J. Combin. **28** (2007), 1131–1143.

[52] D. Hefetz and S. Stich, *On two problems regarding the Hamilton cycle game*, Electron. J. Combin. **16** (1) (2009), publ. R28.

[53] S. Janson, T. Łuczak, and A. Ruciński, *Random graphs*, Wiley, 2000.

[54] J. H. Kim, *The Ramsey number $R(3,t)$ has order of magnitude $t^2/\log t$*, Random Struct. Alg. **7** (1995), 173–207.

[55] F. Knox, *Two constructions relating to conjectures of Beck on positional games*, manuscript, arXiv1212.3345 [math.CO].

[56] M. Krivelevich, *Bounding Ramsey numbers through large deviation inequalities*, Random Struct. Alg. **7** (1995), 145–155.

[57] ———, *The critical bias for the Hamiltonicity game is $(1+o(1))n/\ln n$*, J. Amer. Math. Soc. **24** (2011), 125–131.

[58] M. Krivelevich and B. Sudakov, *Pseudo-random graphs*, In: More sets, graphs and numbers, Bolyai Soc. Math. Stud. Vol. 15 (2006), 199–262

[59] M. Krivelevich and T. Szabó, *Biased positional games and small hypergraphs with large covers*, Electron. J. Combin. **15** (1) (2008), publ. R70.

[60] A. Lehman, *A solution of the Shannon switching game*, J. SIAM **12** (1964), 687–725.

[61] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.

[62] T. Müller and M. Stojaković, *A threshold for the Maker-Breaker clique game*, Random Struct. Alg., to appear.

[63] R. Nenadov, A. Steger, and M. Stojaković, *On the threshold for the Maker-Breaker H-game*, submitted, arXiv1401.4384 [Math.CO].

[64] A. Nilli, *Shelah's proof of the Hales-Jewett theorem*, Mathematics of Ramsey theory, Algorithms Combin., 5, Springer, Berlin, 1990, 150–151.

[65] V. Rödl and A. Ruciński, *Threshold functions for Ramsey properties*, J. Amer. Math. Soc. **8** (1995), 917–942.

[66] V. Rödl and E. Szemerédi, *On size Ramsey numbers of graphs with bounded degree*, Combinatorica **20** (2000), 257–262.

[67] S. Shelah, *Primitive recursive bounds for van der Waerden numbers*. J. Amer. Math. Soc. **1** (1988), 683–697.

[68] J. Spencer, *Asymptotic lower bounds for Ramsey functions*, Discrete Math. **20** (1977), 69–76.

[69] B. Sudakov and V. H. Vu, *Local resilience of graphs*, Random Struct. Alg. **33** (2008), 409–433.

[70] M. Stojaković and T. Szabó, *Positional games on random graphs*, Random Struct. Alg. **26** (2005), 204–223.

[71] L. A. Székely, *On two concepts of discrepancy in a class of combinatorial games*, In: Finite and Infinite Sets, Colloq. Math. Soc. J. Bolyai, Vol. 37, North-Holland, 1984, 679–683.

School of Mathematical Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
E-mail: krivelev@post.tau.ac.il

Hamilton cycles in graphs and hypergraphs: an extremal perspective

Daniela Kühn and Deryk Osthus

Abstract. As one of the most fundamental and well-known NP-complete problems, the Hamilton cycle problem has been the subject of intensive research. Recent developments in the area have highlighted the crucial role played by the notions of expansion and quasi-randomness. These concepts and other recent techniques have led to the solution of several long-standing problems in the area. New aspects have also emerged, such as resilience, robustness and the study of Hamilton cycles in hypergraphs. We survey these developments and highlight open problems, with an emphasis on extremal and probabilistic approaches.

Mathematics Subject Classification (2010). Primary 05C45; Secondary 05C35, 05C65, 05C20.

Keywords. Hamilton cycles, Hamilton decompositions, factorizations, hypergraphs, graph packings and coverings.

1. Introduction

A Hamilton cycle in a graph G is a cycle that contains all the vertices of G. The decision problem of whether a graph contains a Hamilton cycle is among Karp's original list of NP-complete problems [68]. Together with the satisfiability problem SAT and graph colouring, it is probably one of the most well-studied NP-complete problems. The techniques and insights developed for these fundamental problems have also found applications to many more related and seemingly more complex questions.

The main approach to the Hamilton cycle problem has been to prove natural sufficient conditions which are best possible in some sense. This is exemplified by Dirac's classical theorem [36]: *every graph G on $n \geq 3$ vertices whose minimum degree is at least $n/2$ contains a Hamilton cycle.* More generally, one can ask the following 'extremal' question: what value of some easily computable parameter (such as the minimum degree) ensures the existence of a Hamilton cycle? The field has an enormous literature, so we concentrate on recent developments: several long-standing conjectures have recently been solved and new techniques have emerged. In particular, recent trends include the increasing role of probabilistic techniques and viewpoints as well as approaches based on quasi-randomness.

Correspondingly, in this survey we will focus on the following topics: regular graphs and expansion; optimal packings of Hamilton cycles and Hamilton decompositions; random graphs; uniform hypergraphs; counting Hamilton cycles. Notable omissions include the following topics: Hamilton cycles with additional properties (e.g. k-ordered Hamilton cycles);

Proceedings of the International Congress of Mathematicians, Seoul, 2014

pancyclicity; generalized degree conditions (e.g. Ore- and Fan-type conditions); structural constraints (e.g. claw-free and planar graphs) as well as digraphs. Many results in these areas are covered e.g. in the surveys by Gould [51, 52] and Bondy [22]. Digraphs are discussed in [90], though some very recent results on digraphs are also included here.

2. Regular graphs and expansion

2.1. Dense regular graphs. The union of two cliques as well as the complete almost balanced bipartite graph show that the minimum degree bound in Dirac's theorem is best possible. The former graph is disconnected and the latter is not regular. This led Bollobás [16] as well as Häggkvist (see [61]) to (independently) make the following conjecture: *Every t-connected d-regular graph G on n vertices with $d \geq n/(t+1)$ is Hamiltonian.* The case $t = 2$ was settled in the affirmative by Jackson [61].

Theorem 2.1 ([61]). *Every 2-connected d-regular graph on n vertices with $d \geq n/3$ is Hamiltonian.*

However, Jung [67] and independently Jackson, Li and Zhu [63] gave a counterexample to the conjecture for $t \geq 4$. Until recently, the only remaining case $t = 3$ was wide open. Kühn, Lo, Osthus and Staden [86, 87] proved this case for all large n.

Theorem 2.2 ([86, 87]). *There exists an integer n_0 such that every 3-connected d-regular graph on $n \geq n_0$ vertices with $d \geq n/4$ is Hamiltonian.*

The theorem is best possible in the sense that the bound on d cannot be reduced and 3-connectivity cannot be replaced by 2-connectivity. The key to the proof is a structural partition result for dense regular graphs which was proved recently by the same authors [86]: the latter result gives a partition of an arbitrary dense regular graph into a small number of 'robust components', with very few edges between these components. Each robust component is either a 'robust expander' or a 'bipartite robust expander'. Here a graph G is a robust expander if for every set $S \subseteq V(G)$ of 'reasonable size', its neighbourhood $N(S)$ is significantly larger than S, even after some vertices and edges of G are deleted (the precise definition is given in Section 3.4). [86] also contains further applications of this partition result. Similar ideas might also be useful to prove analogues of Theorem 2.1 (say) for directed and oriented graphs (see [90] for such conjectured analogues).

Christofides, Hladký and Máthé [28] used an approach related to that in the proof of Theorem 2.2 to prove the famous 'Lovász conjecture' in the case of dense graphs.

Conjecture 2.3. *Every connected vertex-transitive graph has a Hamilton path.*

In contrast to common belief, Lovász [99] in 1969 actually asked for the construction of a connected vertex-transitive graph containing no Hamilton path. Traditionally however, the Lovász conjecture is always stated in the positive. A related folklore conjecture is the following:

Conjecture 2.4. *Every connected Cayley graph on at least three vertices contains a Hamilton cycle.*

Here a Cayley graph is defined as follows: Let H be a finite group and let $S \subseteq H$ be a subset with $S = S^{-1}$ such that S does not contain the identity. The corresponding Cayley

graph $G(H;S)$ has vertex set equal to H. Two vertices $g, h \in H$ are joined by a edge if and only if there exists $s \in S$ such that $g = sh$. (So every Cayley graph is vertex-transitive.)

Marušic [101] proved Conjecture 2.4 in the case when H is abelian. Alspach [4] conjectured that in this case one even obtains a decomposition of the set of edges of $G(H;S)$ into edge-disjoint Hamilton cycles and at most one perfect matching. For a survey of results on these conjectures, see for example [97].

The following result of Christofides, Hladký and Máthé [28] confirms the 'dense' case of both Conjecture 2.3 and 2.4.

Theorem 2.5 ([28]). *For every $\varepsilon > 0$ there exists an integer n_0 such that every connected vertex-transitive graph on $n \geq n_0$ vertices of degree at least εn contains a Hamilton cycle.*

To prove this result, Christofides, Hladký and Máthé define the notion of 'iron-connectedness' which is related to that of robust expansion and consider a partition of the given vertex-transitive graph into 'iron-connected' components. It would be interesting to find out whether such a partition-based approach can also be extended to sparser graphs.

2.2. Sparse graphs: Toughness and expansion. The extremal examples for Theorem 2.2 indicate that an obstacle to the existence of a Hamilton cycle is the fact that the graph is 'easy to separate' into several pieces. The examples also show that connectivity is not the appropriate notion to use in this context. So a fruitful direction of research has been to study notions which are stronger than connectivity.

One of the most famous conjectures in this direction is the toughness conjecture of Chvátal [30]. It states that if a graph is 'hard to separate' into many pieces, then it contains a Hamilton cycle.

Conjecture 2.6 ([30]). *There is a constant t so that every t-tough graph has a Hamilton cycle.*

Here a graph is *t-tough* if, for every nonempty set $S \subseteq V(G)$, the graph $G - S$ has at most $|S|/t$ components. Trivially, every graph with a Hamilton cycle is 1-tough. Little progress has been made on this conjecture – we only know that if the conjecture holds, then we must have $t \geq 9/4$ [11].

So instead of considering toughness, it has been more rewarding to consider the related (and in some sense stronger) notions of expansion and quasi-randomness. By expansion, we usually mean the following: every small set S of vertices has a neighbourhood $N(S)$ which is large compared to $|S|$ (more formally, $N(S)$ denotes the set of all those vertices which are adjacent to at least one vertex in S). It is well known that expansion is closely linked to eigenvalues of the adjacency matrix: a large eigenvalue gap is equivalent to good expansion properties (in which case we often call such a graph quasi-random). In particular, there is a conjecture of Krivelevich and Sudakov [81] on Hamilton cycles in regular graphs which involves the 'eigenvalue gap'. The conjecture itself would follow from the toughness conjecture.

Conjecture 2.7 ([81]). *There is a constant C such that whenever G is a d-regular graph and the second largest (in absolute value) eigenvalue of the adjacency matrix of G is a most d/C, then G has a Hamilton cycle.*

The best result towards this was proved by Krivelevich and Sudakov [81].

Theorem 2.8 ([81]). *There exists an integer n_0 such that the following holds for all $n \geq n_0$. Suppose that G is a d-regular graph on n vertices and that the second largest (in absolute value) eigenvalue λ of the adjacency matrix of G satisfies $\lambda \leq \frac{(\log\log n)^2}{1000 \log n (\log\log\log n)} d$. Then G has a Hamilton cycle.*

It is known that $\lambda = \Omega(d^{1/2})$ for $d \leq n/2$. So the above result applies for example to quasi-random graphs with $\lambda = \Theta(d^{1/2})$ whose density is polylogarithmic in n, i.e. for quasi-random graphs which are quite sparse.

The proof of Theorem 2.8 makes crucial use of the fact that the eigenvalue condition implies the following: small sets of vertices expand and there are edges between any two large sets of vertices. Hefetz, Krivelevich and Szabó [57] proved the following general result which goes beyond the class of regular graphs and makes explicit use of these conditions.

Theorem 2.9 ([57]). *There exists an integer n_0 such that the following holds for all integers n, d with $n \geq n_0$ and $12 \leq d \leq e^{(\log n)^{1/2}}$. Let $m := \frac{n(\log\log n)\log d}{d \log n \log\log\log n}$. Suppose that G is a graph on n vertices such that $|N(S)| \geq d|S|$ for every $S \subseteq V(G)$ with $|S| \leq m$. Moreover, suppose that there is an edge in G between any two disjoint subsets $A, B \in V(G)$ with $|A|, |B| \geq m/4130$. Then G has a Hamilton cycle.*

The original motivation for this result was a problem on maker-breaker games, but the result also has several other applications, see [57].

3. Packings of Hamilton cycles and decompositions

3.1. Optimal packings of Hamilton cycles in dense graphs. Nash-Williams [107] proved a striking extension of Dirac's theorem: every graph on $n \geq 3$ vertices with minimum degree at least $n/2$ contains not just one but at least $5n/224$ edge-disjoint Hamilton cycles. He conjectured [106, 107] that there should even be $n/4$ of these. This was disproved by Babai (see [106]), who gave a construction showing that one cannot hope for more than (roughly) $n/8$ edge-disjoint Hamilton cycles (see below for details). Nash-Williams subsequently raised the question of finding the best possible bound, which is answered in Corollary 3.2 below.

Recently Csaba, Kühn, Lapinskas, Lo, Osthus and Treglown [31, 32, 82, 85] were able to answer a more general form of this question: *what is the maximum number of edge-disjoint Hamilton cycles one can guarantee in a graph G of minimum degree δ?*

A natural upper bound is obtained by considering the largest degree $\mathrm{reg}_{\text{even}}(G)$ of an even-regular spanning subgraph of G. Let

$$\mathrm{reg}_{\text{even}}(n, \delta) := \min\{\mathrm{reg}_{\text{even}}(G) : |V(G)| = n,\ \delta(G) = \delta\}.$$

Clearly, in general we cannot guarantee more than $\mathrm{reg}_{\text{even}}(n, \delta)/2$ edge-disjoint Hamilton cycles in a graph of order n and minimum degree δ. The next result of Csaba, Kühn, Lapinskas, Lo, Osthus and Treglown [31, 32, 82, 85] shows that this bound is best possible (if $\delta < n/2$, then $\mathrm{reg}_{\text{even}}(n, \delta) = 0$).

Theorem 3.1 ([31, 32, 82, 85]). *There exists an integer n_0 such that every graph G on $n \geq n_0$ vertices contains at least $\mathrm{reg}_{\text{even}}(n, \delta)/2$ edge-disjoint Hamilton cycles.*

The main result in [82] proves Theorem 3.1 unless G is close to one of the two extremal graphs for Dirac's theorem. This allows us in [31, 32, 85] to restrict our attention to the latter situation (i.e. when G is close to the complete balanced bipartite graph or close to the union of two disjoint copies of a clique).

An approximate version of Theorem 3.1 for $\delta \geq n/2 + \varepsilon n$ was obtained earlier by Christofides, Kühn and Osthus [29]. Hartke and Seacrest [56] gave a simpler argument with improved error bounds.

The parameter $\mathrm{reg}_{\mathrm{even}}(n, \delta)$ can be evaluated via Tutte's theorem. It turns out that for $n/2 \leq \delta < n$, we have

$$\mathrm{reg}_{\mathrm{even}}(n, \delta) \sim \frac{\delta + \sqrt{n(2\delta - n)}}{2},$$

(see [29, 55]). In particular, if $\delta \geq n/2$ then $\mathrm{reg}_{\mathrm{even}}(n, \delta) \geq (n-2)/4$. So Theorem 3.1 implies the following explicit bound, which is best possible and answers the above question of Nash-Williams [106, 107].

Corollary 3.2. *There exists an integer n_0 such that every graph G on $n \geq n_0$ vertices with minimum degree $\delta(G) \geq n/2$ contains at least $(n-2)/8$ edge-disjoint Hamilton cycles.*

The following construction (which is based on a construction of Babai, see [106]) shows that the bound in Corollary 3.2 is best possible for $n = 8k + 2$, where $k \in \mathbb{N}$. Consider the graph G consisting of one empty vertex class A of size $4k$, one vertex class B of size $4k+2$ containing a perfect matching and no other edges, and all possible edges between A and B. Thus G has order $n = 8k + 2$ and minimum degree $4k + 1 = n/2$. Any Hamilton cycle in G must contain at least two edges of the perfect matching in B, so G contains at most $\lfloor |B|/4 \rfloor = k = (n-2)/8$ edge-disjoint Hamilton cycles.

A weaker version of Theorem 3.1 for digraphs was proved by Kühn and Osthus in [93]. Ferber, Krivelevich and Sudakov [42] asked whether one can also obtain such a result for oriented graphs.

Recall that Theorem 3.1 is best possible for the class of graphs on n vertices with minimum degree δ. The following conjecture of Kühn, Lapinskas and Osthus [82] would strengthen this in the sense that it would be best possible for every single graph G.

Conjecture 3.3 ([82]). *Suppose that G is a graph on n vertices with minimum degree $\delta(G) \geq n/2$. Then G contains $\mathrm{reg}_{\mathrm{even}}(G)/2$ edge-disjoint Hamilton cycles.*

For $\delta \geq (2 - \sqrt{2} + \varepsilon)n$, this conjecture was proved by Kühn and Osthus [93]. Recently, Ferber, Krivelevich and Sudakov [42] were able to obtain an approximate version of Conjecture 3.3, i.e. a set of $(1-\varepsilon)\mathrm{reg}_{\mathrm{even}}(G)/2$ edge-disjoint Hamilton cycles under the assumption that $\delta(G) \geq (1+\varepsilon)n/2$.

Also, it seems that the following 'dual' version of the problem has not been investigated yet.

Question 3.4. *Given a graph G on n vertices with $\delta(G) > n/2$, how many Hamilton cycles are needed in order to cover all the edges of G?*

A trivial lower bound would be given by $\lceil \Delta(G)/2 \rceil$. However, this cannot always be achieved. Indeed, consider for example the graph G obtained from a complete graph on an odd number n of vertices by deleting an edge xy. Let $\mathcal{C}$ be a collection of Hamilton cycles covering all edges of G. Since both x and y have odd degree, at least one edge at each of x and y has to lie in at least two Hamilton cycles from $\mathcal{C}$. Thus $|\mathcal{C}| > (n-1)/2 = \Delta(G)/2$.

Moreover, it is easy to see that the condition that $\delta > n/2$ in Question 3.4 is needed to ensure that every edge lies in a Hamilton cycle (consider the balanced complete bipartite graph with a single edge in one of the classes). More is known about the probabilistic version of Question 3.4 (see Section 4).

Question 3.4 can be viewed as a restricted version of the following conjecture of Bondy [21], where arbitrary cycle lengths are permitted:

Conjecture 3.5 ([21]). *The edges of every 2-edge-connected graph on n vertices can be covered by at most $2(n-1)/3$ cycles.*

3.2. The Hamilton decomposition and 1-factorization conjectures. Theorem 3.1 shows that for dense graphs the bottleneck for finding many edge-disjoint Hamilton cycles is the densest even-regular spanning subgraph. This makes it natural to consider the class of dense regular graphs. In fact, Nash-Williams [106] suggested that these should even have a Hamilton decomposition.

Here a *Hamilton decomposition* of a graph G consists of a set of edge-disjoint Hamilton cycles covering all edges of G. A natural extension of this to regular graphs G of odd degree is to ask for a decomposition into Hamilton cycles and one perfect matching (i.e. one perfect matching M in G together with a Hamilton decomposition of $G-M$). The most basic result in this direction is Walecki's theorem (see [100]), which dates back to the 19th century:

Theorem 3.6 ([100]). *If n is odd, then the complete graph K_n on n vertices has a Hamilton decomposition. If n is even, then K_n has a decomposition into Hamilton cycles together with a perfect matching.*

The following result of Csaba, Kühn, Lo, Osthus and Treglown [31, 32, 84, 85] generalizes Walecki's theorem to arbitrary regular graphs which are sufficiently dense: it determines the degree threshold for a regular graph to have a Hamilton decomposition. In particular, it solves the above 'Hamilton decomposition conjecture' of Nash-Williams [106] for all large graphs.

Theorem 3.7 ([31, 32, 84, 85]). *There exists an integer n_0 such that the following holds. Let $n, d \in \mathbb{N}$ be such that $n \geq n_0$ and $d \geq \lfloor n/2 \rfloor$. Then every d-regular graph G on n vertices has a decomposition into Hamilton cycles and at most one perfect matching.*

The bound on the degree in Theorem 3.7 is best possible. Indeed, it is easy to see that a smaller degree bound would not even ensure connectivity. Previous results include the following: Nash-Williams [105] showed that the degree bound in Theorem 3.7 guarantees a single Hamilton cycle. Jackson [60] showed that one can guarantee close to $d/2 - n/6$ edge-disjoint Hamilton cycles. Christofides, Kühn and Osthus [29] obtained an approximate decomposition under the assumption that $d \geq n/2 + \varepsilon n$. Under the same assumption, Kühn and Osthus [93] obtained an exact decomposition (as a consequence of Theorem 3.16 below). Note that Conjecture 3.3 would 'almost' imply Theorem 3.7.

Theorem 3.7 is related to the so-called '1-factorization conjecture'. Recall that Vizing's theorem states that for any graph G of maximum degree $\Delta(G)$, the edge-chromatic number $\chi'(G)$ of G is either $\Delta(G)$ or $\Delta(G)+1$. For regular graphs G, $\chi'(G) = \Delta(G)$ is equivalent to the existence of a 1-*factorization*, i.e. of a set of edge-disjoint perfect matchings covering all edges of G. The long-standing 1-factorization conjecture guarantees a 1-factorization in every regular graph of sufficiently high degree. It was first stated explicitly by Chetwynd and

Hilton [26, 27] (who also proved partial results). However, they state that according to Dirac, it was already discussed in the 1950s. The following result of Csaba, Kühn, Lo, Osthus and Treglown [31, 32, 84, 85] confirms this conjecture for sufficiently large graphs.

Theorem 3.8 ([31, 32, 84, 85]). *There exists an n_0 such that the following holds. Let $n, d \in \mathbb{N}$ be such that $n \geq n_0$ is even and $d \geq 2\lceil n/4 \rceil - 1$. Then every d-regular graph G on n vertices has a 1-factorization. Equivalently, $\chi'(G) = d$.*

The bound on the minimum degree in Theorem 3.8 is best possible. Indeed, a smaller bound on d would not even ensure a single perfect matching. To see this, suppose for example that $n = 2 \mod 4$ and consider the graph which is the disjoint union of two cliques of order $n/2$ (which is odd).

Note that Theorem 3.7 does not quite imply Theorem 3.8, as the degree threshold in the former result is slightly higher. The 1-factorization conjecture is a special case of the 'overfull subgraph' conjecture. This would give an even wider class of graphs whose edge-chromatic number equals the maximum degree (see e.g. the monograph [118]).

The best previous result towards the 1-factorization conjecture is due to Perkovic and Reed [109], who proved an approximate version, i.e. they assumed that $d \geq n/2 + \varepsilon n$. This was generalized by Vaughan [121] to multigraphs of bounded multiplicity.

The following 'perfect 1-factorization conjecture' was posed by Kotzig [77] more than fifty years ago at the first international conference devoted to Graph Theory. It combines 1-factorizations and Hamilton decompositions. First note that it is easy to see that the complete graph K_{2n} has a 1-factorization. The 'perfect 1-factorization conjecture' would provide a far-reaching generalization of this fact.

Conjecture 3.9 ([77]). *K_{2n} has a perfect 1-factorization, i.e. a 1-factorization in which any two 1-factors induce a Hamilton cycle.*

The conjecture is known to hold if n or $2n-1$ is a prime, and for several special values of n, but beyond that very little is known. To approach the conjecture it would be interesting to find 1-factorizations so that the number of pairs of 1-factors which induce Hamilton cycles is as large as possible (see e.g. [123]).

Walecki's theorem can also be generalized in another direction: Alspach conjectured that one can decompose the complete graph K_n into cycles of arbitrary length. This was recently confirmed by Bryant, Horsley and Pettersson [24].

Theorem 3.10. *K_n has a decomposition into t cycles of specified lengths $m_1, \dots, m_t$ if and only if n is odd, $3 \leq m_i \leq n$ for $i \leq t$, and $m_1 + \cdots + m_t = \binom{n}{2}$.*

Perhaps it might be possible to prove a probabilistic analogue of this or extend the result to non-complete graphs.

As the final open problem in the area, we turn to a beautiful conjecture of Bermond (see [5]) that the existence of a Hamilton decomposition in a graph is inherited by its line graph (note that an Euler circuit in a graph corresponds to a Hamilton cycle in the line graph).

Conjecture 3.11 (see [5]). *If G has a Hamilton decomposition, then the line graph $L(G)$ of G has a Hamilton decomposition as well.*

Muthusamy and Paulraja [104] proved this conjecture in the case when the number of Hamilton cycles in a Hamilton decomposition of G is even (i.e. when G is d-regular where

$4|d$). They also came quite close to proving it in the remaining case: they showed that if the number of Hamilton cycles in a Hamilton decomposition of G is odd, then $L(G)$ can be decomposed into Hamilton cycles and one 2-factor.

3.3. Kelly's conjecture. Kelly's conjecture (see e.g. [102]) dates back to 1968 and states that every regular tournament has a Hamilton decomposition. So one could view this as an oriented version of Walecki's theorem. Kühn and Osthus [92] recently proved the following result, which shows that Kelly's conjecture is even true if one replaces the class of regular tournaments by that of sufficiently dense regular oriented graphs. (An *oriented graph* G is a directed graph without 2-cycles. G is *d-regular* if all the in- and outdegrees equal d.)

Theorem 3.12 ([92]). *For every $\varepsilon > 0$ there exists an integer n_0 such that every d-regular oriented graph G on $n \geq n_0$ vertices with $d \geq 3n/8 + \varepsilon n$ has a Hamilton decomposition.*

In fact, Kühn and Osthus deduce this result from an even more general result, which involves an expansion condition rather than a degree condition (see Theorem 3.16). It is not clear whether the bound '$3n/8$' is best possible. However, this bound is a natural barrier since the minimum in- and outdegree threshold which guarantees a single Hamilton cycle in an (not necessarily regular) oriented graph is $(3n-4)/8$. As mentioned above, Theorem 3.12 implies Kelly's conjecture for all large tournaments.

Corollary 3.13. *There exists an integer n_0 such that every regular tournament on $n \geq n_0$ vertices has a Hamilton decomposition.*

Kühn and Osthus [93] also used Theorem 3.12 to prove a conjecture of Erdős on optimal packings of Hamilton cycles in random tournaments, which can be viewed as a probabilistic version of Kelly's conjecture:

Theorem 3.14 ([93]). *Let T be a tournament on n vertices which is chosen uniformly at random. Then a.a.s. T contains $\min\{\delta^+(T), \delta^-(T)\}$ edge-disjoint Hamilton cycles.*

(Here we write a.a.s. for 'asymptotically almost surely', see Section 4 for the definition.) The bound is clearly best possible. A similar phenomenon has been shown to occur in the random graph $G_{n,p}$ (see Theorem 4.1).

Jackson [62] posed the following bipartite version of Kelly's conjecture. Here a *bipartite tournament* is an orientation of a complete bipartite graph.

Conjecture 3.15 ([62]). *Every regular bipartite tournament has a Hamilton decomposition.*

It is not even known whether there exists an approximate decomposition, i.e. a set of Hamilton cycles covering almost all the edges of a regular bipartite tournament. Another conjecture related to Kelly's conjecture was posed by Thomassen. The idea is to force many edge-disjoint Hamilton cycles by high connectivity rather than regularity: Thomassen [120] conjectured that for every k there is an integer $f(k)$ so that every strongly $f(k)$-connected tournament contains k edge-disjoint Hamilton cycles. Kühn, Lapinskas, Osthus and Patel [83] proved this by showing that $f(k) = O(k^2(\log k)^2)$ and conjectured that $f(k) = O(k^2)$.

3.4. Robust expansion. As we already indicated in Section 2, there is an intimate connection between expansion and Hamiltonicity. In what follows, we describe a relatively new

'dense' notion of expansion, which has been extremely fruitful in studying not just Hamilton cycles but also Hamilton decompositions and more general subgraph embeddings.

Roughly speaking, this notion of 'robust expansion' is defined as follows: for any set S of vertices, its robust neighbourhood is the set of all those vertices which have many neighbours in S. A graph is a robust expander if for every set S which is not too small and not too large, its robust neighbourhood is at least a little larger than S itself.

More precisely, let $0 < \nu \leq \tau < 1$. Given any graph G on n vertices and $S \subseteq V(G)$, the *ν-robust neighbourhood $RN_{\nu,G}(S)$ of S* is the set of all those vertices x of G which have at least νn neighbours in S. G is called a *robust (ν, τ)-expander* if

$$|RN_{\nu,G}(S)| \geq |S| + \nu n \text{ for all } S \subseteq V(G) \text{ with } \tau n \leq |S| \leq (1-\tau)n.$$

This notion was introduced (for digraphs) by Kühn, Osthus and Treglown [95], who showed that every robustly expanding digraph of linear minimum in- and outdegree contains a Hamilton cycle. Examples of robust expanders include graphs on n vertices with minimum degree at least $n/2 + \varepsilon n$ as well as quasi-random graphs. Kühn and Osthus [92, 93] showed that every sufficiently large regular robust expander of linear degree has a Hamilton decomposition.

Theorem 3.16 ([92, 93]). *For every $\alpha > 0$ there exists $\tau > 0$ such that for all $\nu > 0$ there exists an integer $n_0 = n_0(\alpha, \nu, \tau)$ for which the following holds. Suppose that G is a d-regular graph on $n \geq n_0$ vertices, where $d \geq \alpha n$, and that G is a robust (ν, τ)-expander. Then G has a Hamilton decomposition.*

In [92] they actually proved a version of this for digraphs, which has several applications. (The undirected version is derived in [93].) For example, this digraph version implies the following result.

Theorem 3.17 ([92]). *For every $\varepsilon > 0$ there exists an integer n_0 such that every d-regular digraph G on $n \geq n_0$ vertices with $d \geq (1/2 + \varepsilon)n$ has a Hamilton decomposition.*

Theorem 3.17 is a far-reaching generalization of a result of Tillson, who proved a directed version of Walecki's theorem. Moreover, Theorem 3.17 (which is algorithmic) has an application to finding good tours for the (asymmetric) Traveling Salesman Problem (see [92]).

The main original motivation for these results was to prove Kelly's conjecture for large tournaments: indeed the directed version of Theorem 3.16 easily implies Theorem 3.12.

Theorem 3.16 has numerous further applications apart from Theorems 3.17 and 3.12 (both immediate ones and ones for which it is used as a tool). For example, it is easy to see that for dense graphs, robust expansion is a relaxation of the notion of quasi-randomness. So in particular, Theorem 3.16 implies (for large n) a recent result of Alspach, Bryant and Dyer [6] that every Paley graph has a Hamilton decomposition. Theorem 3.16 is also used in the proof of the Hamilton decomposition conjecture and the 1-factorization conjecture (Theorems 3.7 and 3.8).

The proof of Theorem 3.16 uses an 'approximate' version of the result, which was proved by Osthus and Staden [108] and states that the conditions of the theorem imply the existence of an 'approximate decomposition', i.e. the existence of a set of edge-disjoint Hamilton cycles covering almost all edges of G. (This generalizes an earlier result of Kühn, Osthus and Treglown [96] on approximate Hamilton decompositions of regular tournaments.)

4. Random graphs

Probabilistic versions of the above Hamiltonicity questions have also been studied intensively. As usual, $G_{n,p}$ will denote a binomial random graph on n vertices where every edge is present with probability p (independently from all other edges), and we say that a property of a random graph on n vertices holds *a.a.s.* (asymptotically almost surely) if the probability that it holds tends to 1 as n tends to infinity.

Improving on bounds by several authors, Bollobás [17]; Komlós and Szemerédi [75] as well as Korshunov [76] determined the precise value of p which ensures a Hamilton cycle: if $pn \geq \log n + \log\log n + \omega(n)$, where $\omega(n) \to \infty$ as $n \to \infty$, then a.a.s. $G_{n,p}$ contains a Hamilton cycle. On the other hand, if $pn \leq \log n + \log\log n - \omega(n)$, then a.a.s. $G_{n,p}$ contains an isolated vertex.

One can even obtain a 'hitting time' version of this result in the evolutionary process $G_{n,t}$. For this, let $G_{n,0}$ be the empty graph on n vertices. Consider a random ordering of the edges of K_n. Let $G_{n,t}$ be obtained from $G_{n,t-1}$ by adding the tth edge in the ordering. Ajtai, Kómlos and Szemerédi [1] as well as Bollobás [18] showed that a.a.s. the time t at which $G_{n,t}$ attains minimum degree two is the same as the time at which it first contains a Hamilton cycle.

There are many generalizations and related results. Recently, much attention has focused on optimal packings of edge-disjoint Hamilton cycles and on resilience and robustness, which we will discuss below. However, many intriguing questions remain open.

4.1. Optimal packings of Hamilton cycles. Bollobás and Frieze [20] extended the above hitting time result to packing edge-disjoint Hamilton cycles in random graphs of bounded minimum degree. In particular, this implies the following: suppose that $pn \leq \log n + O(\log\log n)$. Then a.a.s. $G_{n,p}$ has $\lfloor \delta(G_{n,p})/2 \rfloor$ edge-disjoint Hamilton cycles. Frieze and Krivelevich [46] made the striking conjecture that this extends to all p. This has recently been confirmed in a sequence of papers by several teams of authors:

Theorem 4.1. *For any $p = p(n)$, a.a.s. $G_{n,p}$ has $\lfloor \delta(G_{n,p})/2 \rfloor$ edge-disjoint Hamilton cycles.*

We now summarize the results leading to a proof of Theorem 4.1. Here 'exact' refers to a bound of $\lfloor \delta(G_{n,p})/2 \rfloor$, 'approx.' refers to a bound of $(1-\varepsilon)\delta(G_{n,p})/2$, and ε is a positive constant.

authors	range of p	
Ajtai, Komlós, Szemerédi [1]; Bollobás [18]	$\delta(G_{n,p}) = 2$	exact
Bollobás & Frieze [20]	$\delta(G_{n,p})$ bounded	exact
Frieze & Krivelevich [45]	p constant	approx.
Frieze & Krivelevich [46]	$p = \frac{(1+o(1))\log n}{n}$	exact
Knox, Kühn & Osthus [72]	$p \gg \frac{\log n}{n}$	approx.
Ben-Shimon, Krivelevich & Sudakov [12]	$\frac{(1+o(1))\log n}{n} \leq p \leq \frac{1.02\log n}{n}$	exact
Knox, Kühn & Osthus [73]	$\frac{(\log n)^{50}}{n} \leq p \leq 1 - n^{-1/5}$	exact
Krivelevich & Samotij [80]	$\frac{\log n}{n} \leq p \leq n^{-1+\varepsilon}$	exact
Kühn & Osthus [93]	$p \geq 2/3$	exact

In particular, the results in [20, 73, 80, 93] (of which [73, 80] cover the main range) together imply Theorem 4.1.

Glebov, Krivelevich and Szabó [50] were the first to consider the 'dual' version of this problem: *how many Hamilton cycles are needed to cover all the edges of $G_{n,p}$?* Hefetz, Kühn, Lapinskas and Osthus [58] solved this problem for all p that are not too small or too large (based on the main lemma of [73]).

Theorem 4.2 ([58]). *Suppose that $\frac{(\log n)^{117}}{n} \leq p \leq 1-n^{-1/8}$. Then a.a.s. the edges of $G_{n,p}$ can be covered by $\lceil \Delta(G_{n,p})/2 \rceil$ Hamilton cycles.*

It would be interesting to know whether a 'hitting time' version of Theorem 4.2 holds. For this, given a property $\mathcal{P}$, let $t(\mathcal{P})$ denote the *hitting time* of $\mathcal{P}$, i.e. the smallest t so that $G_{n,t}$ has $\mathcal{P}$.

Question 4.3 ([58]). *Let $\mathcal{C}$ denote the property that an optimal covering of a graph G with Hamilton cycles has size $\lceil \Delta(G)/2 \rceil$. Let $\mathcal{H}$ denote the property that a graph G has a Hamilton cycle. Is it true that a.a.s. $t(\mathcal{C}) = t(\mathcal{H})$?*

Note that $\mathcal{C}$ is not monotone. In fact, it is not even the case that for all $t > t(\mathcal{C})$, $G_{n,t}$ a.a.s. has $\mathcal{C}$. Taking $n \geq 5$ odd and $t = \binom{n}{2} - 1$, $G_{n,t}$ is the complete graph with one edge removed – which, as noted at the end of Section 3.1, cannot be covered by $(n-1)/2$ Hamilton cycles. It would be interesting to determine (approximately) the ranges of t such that a.a.s. $G_{n,t}$ has $\mathcal{C}$.

Another natural model of random graphs is of course that of random regular graphs. In this case it seems plausible that we can actually ask for a Hamilton decomposition (and thus obtain an analogue of Theorem 3.7 for sparse random graphs). Indeed, for random regular graphs of bounded degree this was proved by Kim and Wormald [71] and for (quasi-)random regular graphs of linear degree this was proved by Kühn and Osthus [93] (as a consequence of Theorem 3.16). However, the intermediate range remains open:

Conjecture 4.4. *Suppose that $d = d(n) \to \infty$ and $d = o(n)$. Then a.a.s. a random d-regular graph on n vertices has a decomposition into Hamilton cycles and at most one perfect matching.*

So far, not even an approximate version of this is known. One might be able to deduce this from the results in [73].

An analogue of the hitting time result of Bollobás and Frieze [20] for random geometric graphs was proved by Müller, Perez-Gimenez and Wormald [103]. Here the model is that n vertices are placed at random on the unit square and edges are sequentially added in increasing order of edge-length. For fixed $k \geq 1$, they prove that a.a.s. the first edge in the process that creates minimum degree at least k also causes the graph to have $\lfloor k/2 \rfloor$ edge-disjoint Hamilton cycles. The hitting time result for the case $k = 1$ was proved slightly earlier by Balogh, Bollobás, Krivelevich, Müller and Walters [9].

4.2. Resilience. Often one would like to know not just whether some graph G has a property $\mathcal{P}$, but 'how strongly' it has this property. In other words, does G still have property $\mathcal{P}$ if we delete (or add) some edges? Implicitly, variants of this question have been studied for many properties and many classes of graphs. Sudakov and Vu [119] recently initiated the systematic study of this question. In particular, they introduced the notion of *resilience* of a graph with respect to a property $\mathcal{P}$ (below, we assume that $\mathcal{P}$ is monotone increasing, i.e. that $\mathcal{P}$ cannot be destroyed by adding edges): A graph has *local resilience* t with respect to $\mathcal{P}$ if it still has $\mathcal{P}$ whenever one deletes a set of edges such that at each vertex less than t edges

are deleted. A graph has *global resilience* t with respect to $\mathcal{P}$ if it still has $\mathcal{P}$ whenever one deletes less than t edges. Which of these variants is the more natural one to study usually depends on the property $\mathcal{P}$: for 'global' properties such as Hamiltonicity and connectivity the local resilience leads to more interesting results, whereas for 'local' properties such as triangle containment, it makes more sense to study the global resilience. Resilience has been studied intensively for various random graph models (mainly $G_{n,p}$), as it yields natural probabilistic versions of 'classical' theorems. Lee and Sudakov [98] proved a resilience version of Dirac's theorem (which improved previous bounds by several authors):

Theorem 4.5 ([98]). *For any $\varepsilon > 0$ there is a constant C so that the following holds. If $p \geq C \log n/n$ then a.a.s. every subgraph of $G_{n,p}$ with minimum degree at least $(1+\varepsilon)np/2$ contains a Hamilton cycle.*

It is natural to consider more general structures than Hamilton cycles. However, as observed by Huang, Lee and Sudakov [59], there is a limit to what one can ask for in this context: for every $\varepsilon > 0$ there exists p with $0 < p < 1$ such that a.a.s. $G_{n,p}$ contains a subgraph H with minimum degree at least $(1-\varepsilon)np$ and $\Omega(1/p^2)$ vertices that are not contained in a triangle of H.

As an even more informative notion than local resilience, Lee and Sudakov [98] recently suggested a generalization of local resilience which allows a different number of edges to be deleted at different vertices. In other words, in this 'degree sequence resilience' the degree sequence of the deleted graph has to be dominated by the given constraints. In particular, they asked for a resilience version of Chvátal's theorem on Hamilton cycles:

Problem 4.6 ([98]). *Characterize all those sequences $(k_1, \dots, k_n)$ for which $G = G_{n,p}$ a.a.s. has the following property: Let $H \subseteq G$ be such that the degree sequence $(d_1, \dots, d_n)$ of H satisfies $d_i \leq k_i$ for all $i \leq n$. Then $G - H$ has a Hamilton cycle.*

Partial results on this problem were obtained by Ben-Shimon, Krivelevich and Sudakov [12].

4.3. Robust Hamiltonicity. An approach which can be viewed as 'dual' to resilience was taken by Krivelevich, Lee and Sudakov [79]. They proved the following extension of Dirac's theorem, which one can view as a 'robust' version of the theorem.

Theorem 4.7 ([79]). *There exists a constant C such that for $p \geq C \log n/n$ and a graph G on n vertices of minimum degree at least $n/2$, the random subgraph G_p obtained from G by including each edge with probability p is a.a.s. Hamiltonian.*

This theorem gives the correct order of magnitude of the threshold function since if p is a little smaller than $\log n/n$, then the graph G_p a.a.s. has isolated vertices. Also, since there are graphs with minimum degree $n/2 - 1$ which are not even connected, the minimum degree condition cannot be improved. Note that the result can be viewed as an extension of Dirac's theorem since the case $p = 1$ is equivalent to Dirac's theorem.

One can ask similar questions for other (families of) graphs which are known to be Hamiltonian. In particular, a natural question that seems to have been unfairly neglected is that of the Hamiltonicity threshold in random hypercubes. More precisely, given n and p, the random subgraph $Q_{n,p}$ of the n-dimensional cube Q_n is defined as follows: each edge of Q_n is included independently in $Q_{n,p}$ with probability p. Bollobás [19] proved that if $p > 1/2$ is a constant, then a.a.s. $Q_{n,p}$ is connected and has a perfect matching (and actually proved a hitting time version of this result). It seems plausible that a.a.s. $Q_{n,p}$ even contains

a Hamilton cycle. There is no chance for this if $p \leq 1/2$ as there is a significant probability that $Q_{n,p}$ has an isolated vertex in that case.

Conjecture 4.8. *Suppose that $p > 1/2$ is a constant. Then a.a.s. $Q_{n,p}$ has a Hamilton cycle.*

As far as we are aware, the question is still open even if p is any constant close to one. Since Q_n is Hamiltonian, the above conjecture can be viewed as a 'robust' version of this simple fact.

4.4. The Pósa-Seymour conjecture. Surprisingly, a probabilistic analogue of the Pósa-Seymour conjecture is still open. This beautiful generalization of Dirac's theorem states that every graph G on n vertices with minimum degree at least $kn/(k+1)$ contains the kth power of a Hamilton cycle (which is obtained from a Hamilton cycle C by adding edges between any vertices at distance at most k on C). The conjecture was proved for large graphs by Komlós, Sárközy and Szemerédi [74]. For squares of Hamilton cycles (i.e. for $k = 2$) the best current bound in this direction is due to Châu, DeBiasio and Kierstead [25], who proved that in this case the conjecture holds for all graphs on at least $2 \cdot 10^8$ vertices.

A straightforward first moment argument indicates that the threshold for the square of a Hamilton cycle in $G_{n,p}$ should be close to $p = n^{-1/2}$. Note that unlike the deterministic version of the problem, this threshold would be significantly larger than the threshold for a triangle-factor. The latter was determined to be $n^{-2/3}(\log n)^{1/3}$ in a breakthrough by Johansson, Kahn and Vu [66].

Conjecture 4.9 ([91]). *If $p \gg n^{-1/2}$, then a.a.s. $G_{n,p}$ contains the square of a Hamilton cycle.*

When $k \geq 3$, the threshold is $n^{-1/k}$. This follows from a far more general theorem on thresholds for spanning structures in $G_{n,p}$ which was obtained by Riordan [110]. His proof is based on the second moment method. In [91] Kühn and Osthus proved an 'approximate' version of the above conjecture: for any $\varepsilon > 0$, if $p \geq n^{-1/2+\varepsilon}$, then $G_{n,p}$ a.a.s. contains the square of a Hamilton cycle. Their proof is 'combinatorial' in the sense that it uses a version of the absorbing method for random graphs rather than the second moment method. A version of this for quasi-random graphs was proved by Allen, Böttcher, Hàn, Kohayakawa and Person [2]. Their result also extends to kth powers of Hamilton cycles.

In the spirit of Theorem 4.7, one could also ask about a 'robust' version of Conjecture 4.9.

5. Hamilton cycles in uniform hypergraphs

Cycles in hypergraphs have been studied since the 1970s. The first notion of a hypergraph cycle was introduced by Berge [13]. Recently, the much more structured notion of 'ℓ-cycles' has become very popular and has led to very interesting results.

5.1. Dirac-type theorems. To obtain analogues of Dirac's theorem for hypergraphs, we first need to generalize the notions of a cycle and of minimum degree. There are several natural notions available.

A *k-uniform hypergraph* G consists of a set $V(G)$ of vertices and a set $E(G)$ of edges so that each edge of consists of k vertices. Given an integer ℓ with $1 \leq \ell < k$, we say that a k-uniform hypergraph C is an *ℓ-cycle* if there exists a cyclic ordering of the vertices

of C such that every edge of C consists of k consecutive vertices and such that every pair of consecutive edges (in the natural ordering of the edges) intersects in precisely ℓ vertices. So every ℓ-cycle C has $|V(C)|/(k-\ell)$ edges. In particular, $k-\ell$ divides the number of vertices in C. If $\ell = k-1$, then C is called a *tight cycle*, and if $\ell = 1$, then C is called a *loose cycle*. C is a *Hamilton ℓ-cycle* of a k-uniform hypergraph G if $V(C) = V(G)$ and $E(C) \subseteq E(G)$.

More generally, a *Berge cycle* is an alternating sequence $v_1, e_1, v_2, \dots, v_n, e_n$ of distinct vertices v_i and distinct edges e_i so that each e_i contains v_i and v_{i+1}. (Here $v_{n+1} := v_1$, and the edges e_i are also allowed to contain vertices outside $\{v_1, \dots, v_n\}$.) Thus every ℓ-cycle is also a Berge cycle. A Berge cycle $v_1, e_1, v_2, \dots, v_n, e_n$ is a *Hamilton Berge cycle* of a hypergraph G if $V(G) = \{v_1, \dots, v_n\}$ and $e_i \in E(G)$ for each $i \le n$. So a Hamilton Berge cycle of G has $|V(G)|$ edges. Moreover, every tight Hamilton cycle of G is also a Hamilton Berge cycle of G (but this is not true for Hamilton ℓ-cycles with $\ell \le k-2$ as they have $|V(G)|/(k-\ell)$ edges).

We now introduce several notions of minimum degree for a k-uniform hypergraph G. Given a set S of vertices of G, the *degree* $d_G(S)$ of S is the number of all those edges of G which contain S as a subset. The *minimum t-degree* $\delta_t(G)$ of G is then the minimum value of $d_G(S)$ taken over all sets S of t vertices of G. When $t = 1$ we refer to this as the *minimum vertex degree* of G, and when $t = k-1$ we refer to this as the *minimum codegree*.

A Dirac-type theorem for Berge cycles was proved by Bermond, Germa, Heydemann and Sotteau [15]. A Dirac-type theorem for tight Hamilton cycles was proved by Rödl, Ruciński and Szemerédi [113, 114]. (This improved an earlier bound by Katona and Kierstead [69].) Together with the fact that if $(k-\ell)|n$ then any tight cycle contains an ℓ-cycle on the same vertex set (consisting of every $(k-\ell)$th edge), this yields the following result.

Theorem 5.1 ([113, 114]). *For all $k \ge 3$, $1 \le \ell \le k-1$ and any $\varepsilon > 0$ there exists an integer n_0 so that if $n \ge n_0$ and $(k-\ell)|n$ then any k-uniform hypergraph G on n vertices with $\delta_{k-1}(G) \ge \left(\frac{1}{2} + \varepsilon\right) n$ contains a Hamilton ℓ-cycle.*

If $(k-\ell)|k$ and $k|n$ then the above result is asymptotically best possible. Indeed, to see this, note that if the above divisibility conditions hold, then every ℓ-cycle C contains a perfect matching (consisting of every $k/(k-\ell)$th edge of C). On the other hand, it is easy to see that the following parity based construction shows that a minimum codegree of $n/2 - k$ does not ensure a perfect matching: Given a set V of n vertices, let $A \subseteq V$ be a set of vertices such that $|A|$ is odd and $n/2 - 1 \le |A| \le n/2 + 1$. Let G be the k-uniform hypergraph whose edges consists of all those k-element subsets S of V for which $|S \cap A|$ is even.

For $k = 3$, Rödl, Ruciński and Szemerédi [115] were able to prove an exact version of Theorem 5.1 (the threshold in this case is $\lfloor n/2 \rfloor$). The following result of Kühn, Mycroft and Osthus [88] deals with all those cases in which Theorem 5.1 is not asymptotically best possible.

Theorem 5.2 ([88]). *For all $k \ge 3$, $1 \le \ell \le k-1$ with $(k-\ell) \nmid k$ and any $\varepsilon > 0$ there exists an integer n_0 so that if $n \ge n_0$ and $(k-\ell)|n$ then any k-uniform hypergraph G on n vertices with*

$$\delta_{k-1}(G) \ge \left(\frac{1}{\lceil \frac{k}{k-\ell} \rceil (k-\ell)} + \varepsilon \right) n$$

contains a Hamilton ℓ-cycle.

Theorem 5.2 is asymptotically best possible. To see this, let $t := n/(k-\ell)$ and $s := \lceil k/(k-\ell)\rceil$. Fix a set A of $\lceil t/s\rceil - 1$ vertices and consider the k-uniform hypergraph G on n vertices whose hyperedges all have nonempty intersection with A. Then $\delta_{k-1}(G) = |A|$. However, an ℓ-cycle on n vertices has t edges and every vertex on such a cycle lies in at most s edges. So G does not contain an Hamilton ℓ-cycle since A would be a vertex cover for such a cycle and $|A|s < t$.

So the problem of which codegree forces a Hamilton ℓ-cycle is asymptotically solved, though exact versions covering all cases remain a challenging open problem. For $k = 3$ and $\ell = 1$, Czygrinow and Molla [35] were able to prove such an exact version. The following table describes the history of the results leading to the current state of the art.

authors	k	range of ℓ	
Rödl, Ruciński & Szemerédi [113]	$k = 3$	$\ell = 2$	approx.
Kühn & Osthus [89]	$k = 3$	$\ell = 1$	approx.
Rödl, Ruciński & Szemerédi [114]	$k \geq 3$	$\ell = k-1$	approx.
Keevash, Kühn, Mycroft & Osthus [70]	$k \geq 3$	$\ell = 1$	approx.
Hàn & Schacht [53]	$k \geq 3$	$1 \leq \ell < k/2$	approx.
Kühn, Mycroft & Osthus [88]	$k \geq 3$	$1 \leq \ell < k-1,\ (k-\ell) \nmid k$	approx.
Rödl, Ruciński & Szemerédi [115]	$k = 3$	$\ell = 2$	exact
Czygrinow and Molla [35]	$k = 3$	$\ell = 1$	exact

Proving corresponding results for vertex degrees seems to be considerably harder. The following natural conjecture, which was implicitly posed by Rödl and Ruciński [111], is wide open.

Conjecture 5.3 ([111]). *For all integers $k \geq 3$ and all $\varepsilon > 0$ there is an integer n_0 so that the following holds: if G is a k-uniform hypergraph on $n \geq n_0$ vertices with*

$$\delta_1(G) \geq \left(1 - \left(1 - \frac{1}{k}\right)^{k-1} + \varepsilon\right)\binom{n}{k-1},$$

then G contains a tight Hamilton cycle.

This would be asymptotically best possible. Indeed, if $k|n$ then any tight Hamilton cycle contains a perfect matching, and a minimum vertex degree which is slightly smaller than in Conjecture 5.3 would not even guarantee a perfect matching. To see the latter, fix a set A of $n/k - 1$ vertices and consider the k-uniform hypergraph G on n vertices whose hyperedges all have nonempty intersection with A. Then $\delta_1(G) \sim (1-(1-1/k)^{k-1})\binom{n}{k-1}$, but G does not contain a perfect matching.

For general k, Conjecture 5.3 seems currently out of reach – it is even a major open question to determine whether the above degree bound ensures a perfect matching of G. However, it would also be interesting to obtain non-trivial bounds (see e.g. [111]). For $k = 3$ the best current bound towards Conjecture 5.3 was proved by Rödl and Ruciński [112]. They showed that in this case the conjecture holds if $1 - (1-1/3)^2 = 5/9$ is replaced by $(5-\sqrt{5})/3$.

For $k = 3$, Han and Zhao [54] were able to determine the minimum vertex degree which guarantees a loose Hamilton cycle exactly.

Theorem 5.4 ([54]). *There exists an integer n_0 so that the following holds. Suppose that G is a 3-uniform hypergraph on $n \geq n_0$ vertices with $\delta_1(G) \geq \binom{n}{2} - \binom{3n/4}{2} + c$, where n is even, $c = 2$ if $4|n$ and $c = 1$ otherwise. Then G contains a loose Hamilton cycle.*

The bound on the minimum vertex degree is tight: for n of the form $4t+2$, fix a set A of t vertices and consider the k-uniform hypergraph G on n vertices whose hyperedges all have nonempty intersection with A. Buß, Han and Schacht [23] had earlier proved an asymptotic version of this result.

5.2. Hamilton cycles in random hypergraphs. Similarly as in the graph case, it is natural to study Hamiltonicity questions in a probabilistic setting. Let $H^{(k)}_{n,p}$ denote the random k-uniform hypergraph on n vertices where every edge is present with probability p, independently of all other edges. The following result of Dudek, Frieze, Loh and Speiss [39] determines the threshold for the existence of a loose Hamilton cycle in $H^{(k)}_{n,p}$. (In both Theorems 5.5 and 5.6 we only consider those n which satisfy the trivial divisibility condition for the existence of an ℓ-cycle, i.e. that n is a multiple of $k-\ell$.)

Theorem 5.5 ([39]). *Suppose that $k \geq 3$. If $p \gg (\log n)/n^{k-1}$, then a.a.s. $H^{(k)}_{n,p}$ contains a loose Hamilton cycle.*

The logarithmic factor appears due to the 'local' obstruction that a.a.s. $H^{(k)}_{n,p}$ contains isolated vertices below this threshold.

The proof of Theorem 5.5 is 'combinatorial' (in particular, it does not use the second moment method) and builds on earlier results by Frieze [44] as well as Dudek and Frieze [37], which required additional divisibility assumptions. The argument in [39] also uses the celebrated result of Johansson, Kahn and Vu [66] on the threshold for perfect matchings in hypergraphs.

Loose Hamilton cycles in random regular hypergraphs have been considered by Dudek, Frieze, Ruciński and Šileikis [40]. The next result due to Dudek and Frieze [38] concerns precisely those values of k and ℓ not covered by Theorem 5.5. Thus together Theorems 5.5 and 5.6 determine the threshold for the existence of a Hamilton ℓ-cycle in random k-uniform hypergraphs for any given value of k and ℓ.

Theorem 5.6 ([40]).

(i) *For all integers $k > \ell \geq 2$ and fixed $\varepsilon > 0$, if $p = (1-\varepsilon)e^{k-\ell}/n^{k-\ell}$, then a.a.s. $H^{(k)}_{n,p}$ does not contain a Hamilton ℓ-cycle.*

(ii) *If $k > \ell \geq 2$ and $p \gg 1/n^{k-\ell}$, then a.a.s. $H^{(k)}_{n,p}$ contains a Hamilton ℓ-cycle.*

(iii) *For all fixed $\varepsilon > 0$, if $k \geq 4$ and $p = (1+\varepsilon)e/n$, then a.a.s. $H^{(k)}_{n,p}$ contains a tight Hamilton cycle.*

The proof of Theorem 5.6 is based on the second moment method (which seems to fail for Theorem 5.5). An algorithmic proof of (iii) with a weaker threshold of $p \geq n^{-1+\varepsilon}$ was given by Allen, Böttcher, Kohayakawa and Person [3]. Note that, for $k \geq 4$, (i) and (iii) establish a sharp threshold for tight Hamilton cycles, i.e. when $\ell = k-1$. It would be interesting to obtain a sharp threshold for other cases besides those in (iii) and a hitting time result for loose Hamilton cycles.

5.3. Hamilton decompositions. Hypergraph generalisations of Walecki's theorem (Theorem 3.6) have also been investigated. This question was first studied for the notion of a Berge cycle. Let $K_n^{(k)}$ denote the complete k-uniform hypergraph on n vertices. Since every Hamilton Berge cycle of $K_n^{(k)}$ has n edges, a necessary condition for the existence of a decomposition of $K_n^{(k)}$ into Hamilton Berge cycles is that n divides $\binom{n}{k}$. Bermond, Germa, Heydemann and Sotteau [15] conjectured that this condition is also sufficient. For $k = 3$, this conjecture follows by combining the results of Bermond [14] and Verrall [122]. Kühn and Osthus [94] showed that as long as n is not too small, the conjecture holds for $k \geq 4$ as well. So altogether this yields the following result.

Theorem 5.7 ([14, 94, 122]). *Suppose that $3 \leq k < n$, that n divides $\binom{n}{k}$ and, in the case when $k \geq 4$, that $n \geq 30$. Then $K_n^{(k)}$ has a decomposition into Hamilton Berge cycles.*

The following conjecture of Kühn and Osthus [94] would be an analogue of Theorem 5.7 for Hamilton ℓ-cycles.

Conjecture 5.8 ([94]). *For all integers $1 \leq \ell < k$ there exists an integer n_0 such that the following holds for all $n \geq n_0$. Suppose that $k - \ell$ divides n and that $n/(k-\ell)$ divides $\binom{n}{k}$. Then $K_n^{(k)}$ has a decomposition into Hamilton ℓ-cycles.*

To see that the divisibility conditions are necessary, recall that every ℓ-cycle on n vertices contains exactly $n/(k-\ell)$ edges.

The 'tight' case $\ell = k - 1$ of Conjecture 5.8 was already formulated and investigated by Bailey and Stevens [7]. Actually, if n and k are coprime, the case $\ell = k - 1$ already corresponds to a conjecture made independently by Baranyai [10] and Katona concerning 'wreath decompositions'. A k-partite version of the 'tight' case of Conjecture 5.8 was recently proved by Schroeder [117].

Conjecture 5.8 is known to hold 'approximately' (with some additional divisibility conditions on n), i.e. one can find a set of edge-disjoint Hamilton ℓ-cycles which together cover almost all the edges of $K_n^{(k)}$. This is a very special case of results in [8, 47, 48] which together guarantee approximate decompositions of quasi-random uniform hypergraphs into Hamilton ℓ-cycles for $1 \leq \ell < k$ (again, the proofs need n to satisfy additional divisibility constraints).

For example, Frieze, Krivelevich and Loh [48] proved an approximate decomposition result for tight Hamilton cycles in quasi-random 3-uniform hypergraphs, which implies the following result about random hypergraphs.

Theorem 5.9 ([48]). *Suppose that ε, p, n satisfy $\varepsilon^{45} np^{16} \geq (\log n)^{21}$. Then whenever $4|n$, a.a.s. there is a collection of edge-disjoint tight Hamilton cycles of $H_{n,p}^{(3)}$ which cover all but at most an $\varepsilon^{1/15}$-fraction of the edges of $H_{n,p}^{(3)}$.*

The proof proceeds via a reduction to an approximate decomposition result of quasi-random digraphs into Hamilton cycles. This reduction is also the cause for the divisibility requirement. It would be nice to be able to eliminate this requirement. It would also be interesting to know whether the threshold for the existence of an approximate decomposition into Hamilton ℓ-cycles coincides with the threshold for a single Hamilton cycle.

6. Counting Hamilton cycles

In Section 3.1 the aim was to strengthen Dirac's theorem (and other results) by finding many edge-disjoint Hamilton cycles. Similarly, it is natural to omit the condition of edge-disjointness and ask for the total number of Hamilton cycles in a graph. For Dirac graphs (i.e. for graphs on n vertices with minimum degree at least $n/2$), this problem was essentially solved by Cuckler and Kahn [33, 34]. They gave a remarkably elegant formula which asymptotically determines the logarithm of the number of Hamilton cycles.

To state their result, we need the following definitions. For a graph G and edge weighting $x\colon E(G) \to \mathbb{R}^+$, set $h(x) := \sum_{e\in E(G)} x_e \log_2(1/x_e)$, where x_e denotes the weight of the edge e. This is related to the entropy function, except that $\sum_{e\in E(G)} x_e$ is not required to equal 1. We call an edge weighting x a *perfect fractional matching* if $\sum_{e\ni v} x_e = 1$ for each vertex v of G. Finally, let $h(G)$ (the 'entropy' of G) be the maximum of $h(x)$ over all fractional matchings x.

Theorem 6.1 ([33, 34]). *Suppose that G is a graph on n vertices with $\delta(G) \geq n/2$. Then the number of Hamilton cycles in G is*

$$2^{2h(G)-n\log_2 e-o(n)}. \tag{6.1}$$

In particular, the number of Hamilton cycles in G is at least

$$(1-o(1))^n \frac{\delta(G)^n}{n^n} n! \geq \frac{n!}{(2+o(1))^n}. \tag{6.2}$$

(6.2) answers a question of Sárközy, Selkow and Szemerédi [116]. The proof of the lower bound in (6.1) proceeds by considering a random walk which embeds the Hamilton cycles. (6.2) is a consequence of (6.1), but the derivation is nontrivial. (It is easy to derive if G is d-regular, as then setting $x_e := 1/d$ for each edge e of G maximises $h(x)$.) As a general bound on the number of Hamilton cycles in Dirac graphs, (6.2) is best possible (up to lower order terms) – consider for example the complete balanced bipartite graph. In fact, it is an easy consequence of Bregman's theorem on permanents that the first bound in (6.2) is best possible for *any* regular graph.

$h(G)$ can be computed in polynomial time, so one can efficiently obtain a rough estimate for the number of Hamilton cycles in a given Dirac graph. The question of obtaining more precise estimates via randomized algorithms was considered earlier by Dyer, Frieze and Jerrum [41]. For graphs whose minimum degree is at least $n/2 + \varepsilon n$, they obtained a fully polynomial time randomized approximation scheme (FPRAS) for counting the number of Hamilton cycles. (Roughly speaking, an FPRAS is a randomized polynomial time algorithm which gives an answer to a counting problem to within a factor of $1 + o(1)$ with probability $1 - o(1)$.) They asked whether this result can be extended to all Dirac graphs.

Question 6.2 ([41]). *Let $\mathcal{G}$ denote the class of all Dirac graphs, i.e. of all graphs G with minimum degree at least $|V(G)|/2$. Is there an FPRAS for counting the number of Hamilton cycles for all graphs in $\mathcal{G}$?*

Ferber, Krivelevich and Sudakov [42] proved an analogue of (6.2) for oriented graphs whose degree is slightly above the Hamiltonicity threshold.

Counting Hamilton cycles also yields interesting results in the random graph setting. Note that the expected number of Hamilton cycles in $G_{n,p}$ is $p^n(n-1)!/2$. Glebov and

Krivelevich [49] showed that for any p above the Hamiltonicity threshold, a.a.s. the number of Hamilton cycles in $G_{n,p}$ is not too far from this.

Theorem 6.3 ([49]). *Let $p \geq \frac{\log n + \log\log n + \omega(n)}{n}$, where $\omega(n)$ tends to infinity with n. Then a.a.s. the number of Hamilton cycles in $G_{n,p}$ is $(1-o(1))^n p^n n!$.*

For $p = \Omega(n^{-1/2})$, this was already proved by Janson [64], who in fact determined the asymptotic distribution of the number of Hamilton cycles in this range. Surprisingly, his results imply that a.a.s. the number X of Hamilton cycles in $G_{n,p}$ is concentrated below the expected value, i.e. a.a.s. $X/\mathbb{E}(X) \to 0$ for $p \to 0$ (on the other hand, in the $G_{n,m}$ model, X is concentrated at $\mathbb{E}(X)$ in the range when $n^{3/2} \leq m \leq 0.99\binom{n}{2}$). Glebov and Krivelevich [49] also obtained a hitting time version of Theorem 6.3.

Theorem 6.4 ([49]). *In the random graph process $G_{n,t}$, at the very moment the minimum degree becomes two, a.a.s. the number of Hamilton cycles becomes $(1-o(1))^n (\log n/e)^n$.*

Note that at the hitting time t for minimum degree two a.a.s. the edge density p of $G_{n,t}$ is close to $\log n/n$, and so the expression in Theorem 6.4 could also be written as $(1-o(1))^n p^n n!$, which coincides with Theorem 6.3.

A related result of Janson [65] determines the asymptotic distribution of the number of Hamilton cycles in random d-regular graphs for constant $d \geq 3$. Frieze [43] proved a similar formula to that in Theorem 6.3 for dense quasi-random graphs, which was extended to sparse quasi-random graphs by Krivelevich [78].

It turns out that the number of Hamilton cycles in a graph is often closely connected to the number of perfect matchings (indeed the former is always at most the square of the latter). So most of the above papers also contain related results about counting perfect matchings.

References

[1] M. Ajtai, J. Komlós, and E. Szemerédi, *The first occurrence of Hamilton cycles in random graphs*, Ann. Discrete Math. **27** (1985), 173–178.

[2] P. Allen, J. Böttcher, H. Hàn, Y. Kohayakawa, and Y. Person, *Powers of Hamilton cycles in pseudorandom graphs*, preprint.

[3] P. Allen, J. Böttcher, Y. Kohayakawa, and Y. Person, *Tight Hamilton cycles in random hypergraphs*, Random Structures & Algorithms, to appear.

[4] B. Alspach, *Research Problem 59*, Discrete Math. **50** (1984), 115.

[5] B. Alspach, J.C. Bermond, and D. Sotteau, *Decomposition into cycles. I. Hamilton decompositions*, in: in: Cycles and rays (Montreal, PQ, 1987), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. **301**, Kluwer Acad. Publ., Dordrecht (1990), 9–18.

[6] B. Alspach, D. Bryant, and D. Dyer, *Paley graphs have Hamilton decompositions*, Discrete Math. **312** (2012), 113–118.

[7] R. Bailey and B. Stevens, *Hamiltonian decompositions of complete k-uniform hypergraphs*, Discrete Math. **310** (2010), 3088–3095.

[8] D. Bal and A. Frieze, *Packing tight Hamilton cycles in uniform hypergraphs*, SIAM J. Discrete Math. **26** (2012), 435–451.

[9] J. Balogh, B. Bollobás, M. Krivelevich, T. Müller, and M. Walters, *Hamilton cycles in random geometric graphs*, Annals of Applied Probability **21** (2011), 1053–1072.

[10] Zs. Baranyai, *The edge-coloring of complete hypergraphs I*, J. Combin. Theory B **26** (1979), 276–294.

[11] D. Bauer, H.J. Broersma, and H.J. Veldman, *Not every 2-tough graph is Hamiltonian*, Discrete Applied Math. **99** (2000), 317–321.

[12] S. Ben-Shimon, M. Krivelevich, and B. Sudakov, *On the resilience of Hamiltonicity and optimal packing of Hamilton cycles in random graphs*, SIAM J. Discrete Math. **25** (2011), 1176–1193.

[13] C. Berge, *Graphs and Hypergraphs*, North-Holland, Amsterdam, 1979.

[14] J.C. Bermond, *Hamiltonian decompositions of graphs, directed graphs and hypergraphs*, Ann. Discrete Math. **3** (1978), 21–28.

[15] J.C. Bermond, A. Germa, M.C. Heydemann, and D. Sotteau, Hypergraphes hamiltoniens, in *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*, vol. 260 of Colloq. Internat. CNRS, Paris (1973), 39–43.

[16] B. Bollobás, *Extremal Graph Theory*, Academic Press, 1978, p167.

[17] ———, *Almost all regular graphs are Hamiltonian*, European J. Combinatorics **4** (1983), 97–106.

[18] ———, *The evolution of sparse graphs*, Graph theory and combinatorics, Academic Press, London (1984), 35–57.

[19] ———, *Complete matchings in random subgraphs of the cube*, Random Structures & Algorithms **1** (1990), 95–104.

[20] B. Bollobás and A. Frieze, *On matchings and Hamiltonian cycles in random graphs*, Random graphs '83 (Poznan, 1983), North-Holland Math. Stud., 118, North-Holland, Amsterdam (1985), 23–46.

[21] A. Bondy, *Small cycle double covers of graphs. in: Cycles and rays (Montreal, PQ, 1987)*, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. **301**, Kluwer Acad. Publ., Dordrecht (1990), 21–40.

[22] ———, *Basic graph theory: paths and circuits*, in Handbook of Combinatorics, Vol. 1, Elsevier, Amsterdam (1995), 3–110.

[23] E. Buß, H. Hán, and M. Schacht, *Minimum vertex degree conditions for loose Hamilton cycles in 3-uniform hypergraphs*, J. Combin. Theory B, to appear.

[24] D. Bryant, D. Horsley, and W. Pettersson, *Cycle decompositions V: Complete graphs into cycles of arbitrary lengths*, Proc. London Math. Soc., to appear.

[25] P. Châu, L. DeBiasio, and H.A. Kierstead, *Posá's conjecture for graphs of order at least* 2×10^8, Random Structures & Algorithms **39** (2011), 507–525.

[26] A.G. Chetwynd and A.J.W. Hilton, *Regular graphs of high degree are 1-factorizable*, Proc. London Math. Soc. **50** (1985), 193–206.

[27] ———, *1-factorizing regular graphs of high degree—an improved bound*, Discrete Math. **75** (1989), 103–112.

[28] D. Christofides, J. Hladký, and A. Máthé, *Hamilton cycles in dense vertex-transitive graphs*, J. Combin. Theory B, to appear.

[29] D. Christofides, D. Kühn, and D. Osthus, *Edge-disjoint Hamilton cycles in graphs*, J. Combin. Theory B **102** (2012), 1035–1060.

[30] V. Chvátal, *Tough graphs and Hamiltonian circuits*, Discrete Math. **5** (1973), 215–228.

[31] B. Csaba, D. Kühn, A. Lo, D. Osthus, and A. Treglown, *Proof of the 1-factorization and Hamilton decomposition conjectures II: the bipartite case*, preprint.

[32] ———, *Proof of the 1-factorization and Hamilton decomposition conjectures III: approximate decompositions*, preprint.

[33] B. Cuckler and J. Kahn, *Hamiltonian cycles in Dirac graphs*, Combinatorica **29** (2009), 299–326.

[34] ———, *Entropy bounds for perfect matchings and Hamiltonian cycles*, Combinatorica **29** (2009), 327–335.

[35] A. Czygrinow and T. Molla, *Tight co-degree condition for the existence of loose Hamilton cycles in 3-graphs*, preprint.

[36] G. Dirac, *Some theorems on abstract graphs*, Proc. London Math. Soc. **2** (1952), 69–81.

[37] A. Dudek and A. Frieze, *Loose Hamilton cycles in random uniform hypergraphs*, Electronic J. Combinatorics **18** (2011), # P48.

[38] ———, *Tight Hamilton cycles in random uniform hypergraphs*, Random Structures & Algorithms **42** (2013), 374–385.

[39] A. Dudek, A. Frieze, P. Loh, and S. Speiss, *Optimal divisibility conditions for loose Hamilton cycles in random hypergraphs*, Electronic J. Combinatorics **19** (2012), # P44.

[40] A. Dudek, A. Frieze, A. Ruciński, and M. Šileikis, *Loose Hamilton cycles in regular hypergraphs*, Combin. Probab. Comput., to appear.

[41] M. Dyer, A. Frieze, and M. Jerrum, *Approximately counting Hamilton paths and cycles in dense graphs*, SIAM J. Computing **27** (1998), 1262–1272.

[42] A. Ferber, M. Krivelevich, and B. Sudakov, *Counting and packing Hamilton cycles in dense graphs and oriented graphs*, preprint.

[43] A. Frieze, *On the number of perfect matchings and Hamilton cycles in ε-regular non-bipartite graphs*, Electronic J. Combinatorics **7** (2000), # R57.

[44] ———, *Loose Hamilton cycles in random 3-uniform hypergraphs*, Electronic J. Combinatorics **17** (2010), # N28.

[45] A. Frieze and M. Krivelevich, *On packing Hamilton cycles in ε-regular graphs*, J. Combin. Theory B **94** (2005), 159–172.

[46] ———, *On two Hamilton cycle problems in random graphs*, Israel J. Math. **166** (2008), 221–234.

[47] ———, *Packing Hamilton cycles in random and pseudo-random hypergraphs*, Random Structures & Algorithms **41** (2012), 1–22.

[48] A. Frieze, M. Krivelevich, and P. Loh, *Packing tight Hamilton cycles in 3-uniform hypergraphs*, Random Structures & Algorithms **40** (2012), 269–300.

[49] R. Glebov and M. Krivelevich, *On the number of Hamilton cycles in sparse random graphs*, SIAM J. Discrete Math. **27** (2013), 27–42.

[50] R. Glebov, M. Krivelevich, and T. Szabó, *On covering expander graphs by Hamilton cycles*, Random Structures & Algorithms (to appear).

[51] R. Gould, *Advances on the Hamiltonian problem: a survey*, Graphs and Combinatorics **19** (2003), 7–52.

[52] ———, *Recent advances on the Hamiltonian problem: Survey III*, Graphs and Combinatorics **30** (2014), 1–46.

[53] H. Hàn and M. Schacht, *Dirac-type results for loose Hamilton cycles in uniform hypergraphs*, J. Combin. Theory B **100** (2010), 332–346.

[54] J. Han and Y. Zhao, *Minimum vertex degree threshold for loose Hamilton cycles in 3-uniform hypergraphs*, preprint.

[55] S.G. Hartke, R. Martin, and T. Seacrest, *Relating minimum degree and the existence of a k-factor*, research manuscript.

[56] S.G. Hartke and T. Seacrest, *Random partitions and edge-disjoint Hamiltonian cycles*, preprint.

[57] D. Hefetz, M. Krivelevich, and T. Szabó, *Hamilton cycles in highly connected and expanding graphs*, Combinatorica **29** (2009), 547–568.

[58] D. Hefetz, D. Kühn, J. Lapinskas, and D. Osthus, *Optimal covers with Hamilton cycles in random graphs*, Combinatorica, to appear.

[59] H. Huang, C. Lee, and B. Sudakov, *Bandwidth theorem for random graphs*, J. Combin. Theory B **102** (2012), 14–37.

[60] B. Jackson, *Edge-disjoint Hamilton cycles in regular graphs of large degree*, J. London Math. Soc. **19** (1979), 13–16.

[61] _____, *Hamilton cycles in regular 2-connected graphs*, J. Combin. Theory B **29** (1980), 27–46.

[62] _____, *Long paths and cycles in oriented graphs*, J. Graph Theory **5** (1981), 245–252.

[63] _____, H. Li and Y. Zhu, *Dominating cycles in regular 3-connected graphs*, Discrete Math. **102** (1991), 163–176.

[64] S. Janson, The numbers of spanning trees, Hamilton cycles and perfect matchings in a random graph, *Combin. Probab. Comput.* **3** (1994), 97–126.

[65] _____, *Random regular graphs: asymptotic distributions and contiguity*, Combin. Probab. Comput. **4** (1995), 369–405.

[66] A. Johansson, J. Kahn, and V. Vu, *Factors in random graphs*, Random Structures & Algorithms **33** (2008), 1–28.

[67] H.A. Jung, *Longest circuits in 3-connected graphs*, Finite and infinite sets, Vol I, II, Colloq. Math. Soc. János Bolyai **37** (1984), 403–438.

[68] R. Karp, *Reducibility among combinatorial problems, Complexity of computer computations*, Plenum, New York, 1972, pp. 85–103.

[69] G.Y. Katona and H.A. Kierstead, *Hamiltonian chains in hypergraphs*, J. Graph Theory **30** (1999), 205–212.

[70] P. Keevash, D. Kühn, R. Mycroft, and D. Osthus, *Loose Hamilton cycles in hypergraphs*, Discrete Math. **311** (2011), 544–559.

[71] J.H. Kim and N. Wormald, *Random matchings which induce Hamilton cycles and Hamiltonian decompositions of random regular graphs*, J. Combin. Theory B **81** (2001), 20–44.

[72] F. Knox, D. Kühn, and D. Osthus, *Approximate Hamilton decompositions of random graphs*, Random Structures & Algorithms **40** (2012), 133–149.

[73] _____, *Edge-disjoint Hamilton cycles in random graphs*, Random Structures & Algorithms, to appear.

[74] J. Komlós, G. N. Sárközy, and E. Szemerédi, *Proof of the Seymour conjecture for large graphs*, Ann. Combin. **2** (1998), 43–60.

[75] J. Komlós and E. Szemerédi, *Limit distribution for the existence of Hamilton cycles in random graphs*, Discrete Math. **43** (1983), 55–63.

[76] A.D. Korshunov, *Solution of a problem of P. Erdős and A. Rényi on Hamiltonian cycles in non-oriented graphs*, Diskret. Analiz. **31** (1977), 17–56 (in Russian).

[77] A. Kotzig, *Hamilton graphs and Hamilton circuits*, Theory of Graphs and its Applications, Proceedings of the Symposium of Smolenice, 1963, Nakl. ČSAV Praha (1964), 62–82.

[78] M. Krivelevich, *On the number of Hamilton cycles in pseudo-random graphs*, Electronic J. Combinatorics **19** (2012), #P25.

[79] M. Krivelevich, C. Lee, and B. Sudakov, *Robust Hamiltonicity of Dirac graphs*, Transactions Amer. Math. Soc. **366** (2014), 3095–3130.

[80] M. Krivelevich and W. Samotij, *Optimal packings of Hamilton cycles in sparse random graphs*, SIAM J. Discrete Math. **26** (2012), 964–982.

[81] M. Krivelevich and B. Sudakov, *Sparse pseudo-random graphs are Hamiltonian*, J. Graph Theory **42** (2003), 17–33.

[82] D. Kühn, J. Lapinskas, and D. Osthus, *Optimal packings of Hamilton cycles in graphs of high minimum degree*, Combin. Probab. Comput. **22** (2013), 394–416.

[83] D. Kühn, J. Lapinskas, D. Osthus, and V. Patel, *Proof of a conjecture of Thomassen on Hamilton cycles in highly connected tournaments*, Proc. London Math. Soc, to appear.

[84] D. Kühn, A. Lo, and D. Osthus, *Proof of the* 1*-factorization and Hamilton decomposition conjectures IV: exceptional systems for the two cliques case*, preprint.

[85] D. Kühn, A. Lo, D. Osthus, and A. Treglown, *Proof of the* 1*-factorization conjecture I: the* 2*-clique case*, preprint.

[86] D. Kühn, A. Lo, D. Osthus, and K. Staden, *The robust component structure of dense regular graphs and applications*, preprint.

[87] ———, *Solution to a problem of Bollobás and Häggkvist on Hamilton cycles in regular graphs*, preprint.

[88] D. Kühn, R. Mycroft, and D. Osthus, *Hamilton ℓ-cycles in uniform hypergraphs*, J. Combin. Theory A **117** (2010), 910–927.

[89] D. Kühn and D. Osthus, *Loose Hamilton cycles in 3-uniform hypergraphs of high minimum degree*, J. Combin. Theory B **96** (2006), 767–821.

[90] ———, *A survey on Hamilton cycles in directed graphs*, European J. Combinatorics **33** (2012), 750–766.

[91] D. Kühn and D. Osthus, *On Posa's conjecture for random graphs*, SIAM J. Discrete Math. **26** (2012), 1440–1457.

[92] ———, *Hamilton decompositions of regular expanders: a proof of Kelly's conjecture for large tournaments*, Adv. in Math. **237** (2013), 62–146.

[93] ———, *Hamilton decompositions of regular expanders: applications*, J. Combin. Theory B **104** (2014), 1–27.

[94] ———, *Decompositions of complete uniform hypergraphs into Hamilton Berge cycles*, J. Combin. Theory A **126** (2014), 128–135.

[95] D. Kühn, D. Osthus, and A. Treglown, *Hamiltonian degree sequences in digraphs*, J. Combin. Theory B **100** (2010), 367–380.

[96] ———, *Hamilton decompositions of regular tournaments*, Proc. London Math. Soc. **101** (2010), 303–335.

[97] K. Kutnar and D. Marušič, *Hamilton cycles and paths in vertex-transitive graphs – current directions*, Discrete Math. **309** (2009), 5491–5500.

[98] C. Lee and B. Sudakov, *Dirac's theorem for random graphs*, Random Structures & Algorithms **41** (2012), 293–305.

[99] L. Lovász. Problem 11. In *Combinatorial Structures and their Applications*, Proceedings of the Calgary International Conference on Combinatorial Structures and their Applications, R. Guy, H. Hanani, N. Sauer, and J. Schönheim, editors, Gordon and Breach Science Publishers, New York, 1970.

[100] E. Lucas, *Récréations Mathématiques*, Vol. 2, Gautheir-Villars, 1892.

[101] D. Marušic, *Hamiltonian circuits in Cayley graphs*, Discrete Math. **46** (1983), 49–54.

[102] J.W. Moon, *Topics on tournaments*, Holt, Rinehart and Winston, New York, 1968.

[103] T. Müller, X. Pérez-Gimenez, and N. Wormald, *Disjoint Hamilton cycles in the random geometric graph*, J. Graph Theory **68** (2011), 299–322.

[104] A. Muthusamy and P. Paulraja, *Hamilton cycle decomposition of line graphs and a conjecture of Bermond*, J. Combin. Theory B **64** (1995), 1–16.

[105] C.St.J.A. Nash-Williams, *Valency sequences which force graphs to have Hamiltonian circuits*, University of Waterloo Research Report, Waterloo, Ontario, 1969.

[106] ———, *Hamiltonian lines in graphs whose vertices have sufficiently large valencies*, in Combinatorial theory and its applications, III (Proc. Colloq., Balatonfüred, 1969), North-Holland, Amsterdam (1970), 813–819.

[107] ———, *Edge-disjoint Hamiltonian circuits in graphs with vertices of large valency*, in Studies in Pure Mathematics (Presented to Richard Rado), Academic Press, London (1971), 157–183.

[108] D. Osthus and K. Staden, *Approximate Hamilton decompositions of regular robustly expanding digraphs*, SIAM J. Discrete Math. **27** (2013), 1372–1409.

[109] L. Perkovic and B. Reed, *Edge coloring regular graphs of high degree*, Discrete Math. **165/166** (1997), 567–578.

[110] O. Riordan, *Spanning subgraphs of random graphs*, Combin. Probab. Comput. **9** (2000), 125–148.

[111] V. Rödl and A. Ruciński, *Dirac-type questions for hypergraphs – a survey (or more problems for Endre to solve)*, in: An Irregular Mind (Szemerédi is 70), Bolyai Soc. Math. Studies **21** (2010), 561–590.

[112] ———, *Families of triples with high minimum degree are Hamiltonian*, Discuss. Math. – Graph Th. **34** (2014) 363–383.

[113] V. Rödl, A. Ruciński, and E. Szemerédi, *A Dirac-type theorem for 3-uniform hypergraphs*, Combin. Probab. Comput. **15** (2006), 229–251.

[114] ———, *An approximate Dirac-type theorem for k-uniform hypergraphs*, Combinatorica **28** (2008), 229–260.

[115] ———, *Dirac-type conditions for hamiltonian paths and cycles in 3-uniform hypergraphs*, Adv. in Math. **227** (2011), 1225–1299.

[116] G. Sárközy, S. Selkow, and E. Szemerédi, *On the number of Hamiltonian cycles in Dirac graphs*, Discrete Math. **265** (2003), 237–250.

[117] M.W. Schroeder, *On Hamilton cycle decompositions of r-uniform r-partite hypergraphs*, Discrete Math. **315/316** (2014), 1–8.

[118] M. Stiebitz, D. Scheide, B. Toft, and L.M. Favrholdt, *Graph Edge Coloring: Vizing's Theorem and Goldberg's Conjecture*, Wiley, 2012.

[119] B. Sudakov and V. Vu, *Local resilience of graphs*, Random Structures & Algorithms **33** (2008), 409–433.

[120] C. Thomassen, *Edge-disjoint Hamiltonian paths and cycles in tournaments*, Proc. London Math. Soc. **45** (1982), 151–168.

[121] E. Vaughan, *An asymptotic version of the multigraph* 1*-factorization conjecture*, J. Graph Theory **72** (2013), 19–29.

[122] H. Verrall, *Hamilton decompositions of complete 3-uniform hypergraphs*, Discrete Math. **132** (1994), 333–348.

[123] D. Wagner, *On the perfect* 1*-factorization conjecture*, Discrete Math. **104** (1992), 211–215.

School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
E-mail: d.kuhn@bham.ac.uk

School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
E-mail: d.osthus@bham.ac.uk

Random planar graphs and beyond

Marc Noy

Abstract. We survey several results on the enumeration of planar graphs and on properties of random planar graphs. This includes basic parameters, such as the number of edges and the number of connected components, and extremal parameters such as the size of the largest component, the diameter and the maximum degree. We discuss extensions to graphs on surfaces and to classes of graphs closed under minors. Analytic methods provide very precise results for random planar graphs. The results for general minor-closed classes are less precise but hold with wider generality.

Mathematics Subject Classification (2010). Primary 05A16; Secondary 05B80.

Keywords. Asymptotic enumeration, random graphs, planar graphs, graph minors.

1. Introduction

The theory of random graphs, initiated by Erdős and Rényi [34] in the early 1960s, has become one of the main areas of research in combinatorics [13, 46]. The model studied originally was the class $G(n, M)$ of graphs with n labelled vertices and M edges, equipped with the uniform distribution. Closely related is the binomial model $G(n, p)$, in which every possible edge between two vertices is selected independently with probability p. The two models are very similar if $p\binom{n}{2}$ is close to M. The advantage of the $G(n, p)$ model is the key property of *independence*, which allows to compute probabilities of basic events exactly, and to determine precise thresholds for basic properties such as being acyclic, connected or Hamiltonian. For instance, the probability that three given vertices span a triangle is exactly p^3, and the probability that a given vertex is isolated is $(1 - p)^{n-1}$.

Things become more difficult if we want to analyze random graphs subject to a global condition, such as being regular, planar or triangle-free. Consider the property of being triangle-free: we cannot select edges independently of each other, since once some edges are selected, other edges are forbidden because they would create triangles. How does one proceed in these cases? Simplifying we can say that there are two ways for analyzing random graphs from a constrained class of graphs: either finding a simpler model that is close enough to the class, or counting graphs in the class, or a combination of both. The first method is well exemplified by the class of regular graphs. In the pairing model for d-regular graphs there are n vertices, each of them equipped with d half-edges. A random pairing of the dn half-edges produces a random d-regular multigraph. Probabilities of elementary events can be computed reasonably well, including the probability that the resulting graph is simple. This allows to obtain precise estimates on the number of regular graphs and has led to a rich theory of random regular graphs [74].

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Another example is the class of triangle-free graphs. It was proved in [33] that, as the number of vertices goes to infinity, almost all triangle-free graph are bipartite. Random bipartite graphs can be analyzed in a model very similar to the $G(n,p)$ model, where again we have independence, and this provides a suitable model for triangle-free graphs. More generally, almost all graphs not containing the complete graph K_t as a subgraph are $(t-1)$-partite [49], and again we have a model similar to $G(n,p)$. Even more generally, if H is a graph with the property that there exists an edge e such that $\chi(H-e) < \chi(H)$ (here χ denotes the chromatic number) and $t = \chi(H) \geq 3$, then almost every graph not containing H as a subgraph is $(t-1)$-partite [67]. These are important examples of monotone classes. A class of graphs is *monotone* if it is closed under taking subgraphs, and it is *hereditary* if it is closed under taking induced subgraphs. Much work has been done on estimating the growth rate of monotone and hereditary classes and on analyzing random graphs from these classes. This is an active area of research closely related to extremal graph theory [14].

The foremost example of the second method for analyzing random graphs, based on counting, is the class of trees. We know how to count trees very precisely (whether labelled or unlabeled, rooted or unrooted) and we also know how to count trees, for instance, with given degrees or with given height. Thus we can analyze random variables like the number of leaves or the height in random trees. Trees are fundamental objects in computer science and powerful methods have been developed for analyzing them. The main tools in this area are generating functions and analytic methods for deriving asymptotic estimates. We enter here the realm of *analytic combinatorics*, as developed by Flajolet and Sedgewick [35]; see also [24] for many aspects of random trees.

The key property that allows us to count trees is that they admit a simple combinatorial decomposition. A rooted tree can be decomposed uniquely into the root and a collection (ordered or not) of subtrees attached to the root. This decomposition translates into equations for the corresponding generating functions, and we are in a situation to apply the methods of analytic combinatorics. Many other combinatorial objects can be decomposed according to simple schemes. This includes the class of planar *maps*. A map is a connected planar multigraph (loops and multiple edges allowed) with a fixed embedding in the plane. In the 1960s Tutte, motivated by the Four Colour Problem, created the theory of map enumeration. He realized that maps admit recursive decompositions, implying algebraic equations for the associated generating functions. He found *exact* formulas for the number of various classes of rooted maps (to be defined later) with given number of edges. For instance, Tutte showed [73] that the number of rooted maps with n edges equals

$$\frac{2 \cdot 3^n (2n)!}{n!(n+2)!}. \tag{1.1}$$

This formula and similar ones were later explained more combinatorially, using bijections with certain classes of enriched trees [70]. As we discuss later, these bijections have powerful implications on the structure of random maps.

It took time to realize that the theory of map enumeration could be used to count planar graphs *without* an embedding. This was done first for 2-connected planar graphs by Bender, Gao and Wormald [9], using the enumeration of 3-connected planar maps and Whitney's theorem, namely that a 3-connected planar graph has a unique embedding in the sphere up to homeomorphism. Soon after that the analysis was extended to arbitrary planar graphs by Giménez and Noy [42]. They provided a precise estimate for the number G_n of planar

graphs with n labelled vertices of the form

$$G_n \sim c\, n^{-7/2} \gamma^n n!, \tag{1.2}$$

where $\gamma \approx 27.2269$ is a well-defined constant, known as the growth constant of planar graphs. This opened the way to the fine analysis of random planar graphs. In the same work [42] it was proved that the number of connected components in a random planar graph follows asymptotically a Poisson distribution plus 1 and that the number of edges is asymptotically Gaussian with linear mean and variance. This is developed in Section 3 for the more basic parameters, and in Section 4 for more advanced extremal parameters, such as the diameter, the maximum vertex degree or the size of the largest block.

The next step was to enumerate graphs that can be embedded in a fixed surface S, orientable or not. McDiarmid [55] showed first that the growth constant for graphs embeddable in a surface does not depend on the surface (a result already known for maps) and is equal to γ. Soon after that the enumeration of graphs on surfaces was completed independently in [7] and [18]. It was shown that the number of labelled graphs with n vertices that can be embedded in the orientable surface of genus g is asymptotically

$$c_g\, n^{5(g-1)/2-1} \gamma^n n!. \tag{1.3}$$

We see that only the subexponential term depend on the genus. It is worth remarking that, unlike the planar case, the counting series of graphs in a surface is not computed exactly but rather sandwiched coefficient-wise between two computable series with the same leading asymptotic terms (more details in Section 5). In addition, it was shown [18] that basic parameters, such as the number of components, the number of edges, or the sizes of the largest component and the largest block, have the same asymptotic distribution as for planar graphs. All these results hold as well for graphs on the non-orientable surface of genus h, in which case the subexponential term in the asymptotics is $n^{5(h-2)/4-1}$.

Graphs on surfaces are strongly related to graph minors. A graph H is a minor of G if H can be obtained from a subgraph of G by contracting edges. A class of graphs is *minor-closed* if it is closed under taking minors. A basic example is the class of planar graphs and, more generally, the class of graphs embeddable in a fixed surface. Other interesting minor-closed classes are series-parallel graphs, ΔY-reducible graphs and graphs with bounded tree-width. Only in a few cases we have access to the counting generating functions, allowing for a precise analysis as in the case of planar graphs or graphs on surfaces. However one can use combinatorial arguments to prove relevant results on random graphs from a minor-closed class. This program has been carried out mostly by McDiarmid and his coauthors. The results are less precise than those obtained using generating functions and analytic methods, but apply to more general situations. One example can illustrate this: the class $\mathcal{K}$ of graphs not containing K_5 as a minor is an interesting class (studied by Wagner, motivated by the Four Colour Problem) containing the class of planar graphs. We do not know how to compute the counting generating function for the class $\mathcal{K}$, but from the results in [56] it follows that the number of components converges to a Poisson law plus 1 and that the expected number of vertices in the largest component is $n - c$ for some constant c. These remarkable results apply to any minor-closed class subject to mild hypothesis, as discussed in Section 6.

There is one general situation where analytic methods still apply, namely when the class of graphs is *subcritical*. This is a technical condition defined in terms of the singularities of the generating functions, but combinatorially it can be interpreted as the fact that the class

contains 'relatively few' 3-connected graphs. This category includes forests, outerplanar graphs, series-parallel graphs and related classes of graphs. Graphs in these classes have typically a tree-like structure and in fact share several properties with trees. The analysis of subcritical classes uses general tools from analytic combinatorics [25, 44] and will be reviewed in Section 7.

We conclude the paper with some remarks and open problems. In the rest of the paper, unless mentioned otherwise, all graphs are labelled and n denotes the number of vertices. For the generating functions that will appear, variable x is associated to vertices and variable y to edges. For maps, n denotes the number of edges and z is the variable associated to edges.

2. Planar maps and graphs

Let us go back to Tutte and the enumeration of planar maps. Rooted trees are easier to enumerate than unrooted ones, since the root vertex gives a starting point for the combinatorial decomposition. In the same way Tutte decided to root maps: an edge (not a vertex) is selected and given an orientation. Let $\mathcal{M}$ be a planar map and let e be its root edge. Tutte's analysis distinguished two cases, depending on whether $\mathcal{M} - e$ is connected or not. In order to keep control of the decomposition, he had to consider the number $M_{n,k}$ of maps with n edges in which the root face (the one to the right of the oriented root edge) has degree k. Analyzing the combinatorial decomposition of maps resulting by removing the root edge, he showed that the generating function $M(z,u) = \sum_{n,k} M_{n,k} u^k z^n$ satisfies the equation

$$M(z,u) = 1 + zu^2 M(z,u)^2 + uz \frac{uM(z,u) - M(z,1)}{u-1}. \tag{2.1}$$

This is a quadratic equation in $M(z,u)$, but we cannot solve it directly since it contains the series $M(z,1)$, which is not independent of the unknown. In order to solve it, Tutte devised what is now known as the *quadratic method*. This is similar to the well-known *kernel method*, but applied to quadratic instead of linear equations. He proved that

$$M(z,1) = \frac{18z - 1 + (1-12z)^{3/2}}{54z^2},$$

and from here the expression in (1.1) follows easily.

Additional techniques allowed Tutte to enumerate various classes of maps. For instance, in order to count bipartite maps it is enough to restrict the degrees of the faces to be even. Eulerian maps are then enumerated by duality. In his seminal paper [73], Tutte also counted maps according to their connectivity. The unique decomposition of connected graphs into 2-connected and 3-connected components allows us to link the generating functions of maps with given connectivity. If $B(z) = \sum B_n z^n$ and $T(z) = \sum T_n z^n$ are, respectively, the generating functions of 2-connected and 3-connected maps (counted according to the number of edges) then the recursive decomposition of a map into its blocks gives

$$M(z) = 1 + B(zM(z)^2). \tag{2.2}$$

If now $h(z)$ is the functional inverse of $(B(z) - 2z)/z$ then the decomposition of a 2-

connected map into its 3-connected components gives

$$T(z) = z^2 - \frac{2z^3}{1+z} - zh(z). \tag{2.3}$$

These equations provide all the information needed. For instance, together with (1.1), it is easy to derive the asymptotic estimates

$$M_n \sim c_M n^{-5/2} 12^n, \qquad B_n \sim c_B n^{-5/2} \left(\frac{27}{4}\right)^n, \qquad T_n \sim c_T n^{-5/2} 4^n,$$

for suitable constants c_M, c_B, c_T.

The enumeration of 3-connected maps is particularly interesting since, by the classical theorem of Steinitz (1922), they correspond precisely to the graphs of convex polytopes in $\mathbb{R}^3$. Another reason of interest is that 3-connected planar graphs have a unique embedding in the sphere, a classical result due to Whitney (1933). It follows that there is a one-to-one correspondence between 3-connected planar maps and 3-connected planar graphs. This leads directly to the enumeration of 3-connected labelled planar graphs in which an edge is distinguished and given a direction, corresponding to the root of the associated map. A key feature is that rooted maps have no non-trivial automorphisms, so that all vertices, edges and faces are distinguishable. We can then give them labels and turn a rooted map into a labelled graph. This was first made explicit in [9]. Now, using again Tutte's decomposition of 2-connected graphs into 3-connected components, but in the reverse direction, it is possible to enumerate 2-connected planar graphs. Let us explain how.

In what follows generating functions for graphs are of the exponential type (whereas for maps are ordinary), and variable x marks vertices and y marks edges. Let $T(x,y)$ be the generating function of 3-connected maps, and $B(x,y)$ that of 2-connected planar graphs. Closely related to $B(x,y)$ is the generating function $D(x,y)$ of 'networks', which are 2-connected graphs rooted at a directed edge (which may be deleted or not) and whose endpoints are not labelled. Then $D(x,y)$ is related to $B(x,y)$ through (see [43] for details)

$$2(1+y)\frac{\partial B}{\partial y}(x,y) = x^2(1+D(x,y)), \tag{2.4}$$

and $D(x,y)$ satisfies the equation

$$D(x,y) = (1+y)\exp\left(\frac{xD(x,y)^2}{1+xD(x,y)} + \frac{2}{x^2}\frac{\partial T}{\partial y}(x, D(x,y))\right) - 1. \tag{2.5}$$

The former two equations are essentially the equivalent of (2.3) for graphs. They are more involved because there are two variables and several derivatives, but it is just Tutte's decomposition applied in the reverse direction: from the knowledge of T we have access to D, hence to B. From here it was shown [9] that the number of 2-connected planar graphs grows like

$$c_B n^{-7/2} (\gamma_B)^n n!,$$

where $\gamma_B \approx 26.18$. Observe that the polynomial growth is $n^{-7/2}$ instead of $n^{-5/2}$, the reason being that maps are rooted and introduce an extra linear factor. This was a major step since little was known on counting planar graphs, as opposed to the rich theory of counting planar maps created by Tutte and greatly extended later on.

It remained to count connected planar graphs using the decomposition of connected graphs into 2-connected components, and then to count planar graphs in general. Let $C(x,y)$ and $G(x,y)$ be the generating functions of connected and arbitrary planar graphs, and let $C^\bullet(x,y) = x\frac{\partial C}{\partial x}(x,y)$ be that of rooted connected graphs, where a vertex is distinguished as the root. The recursive decomposition of a graph into its blocks implies the equation

$$C^\bullet(x,y) = \exp\left(\frac{\partial B}{\partial x}(xC^\bullet(x,y),y)\right). \tag{2.6}$$

The former equation is the analog for graphs of (2.2). And the decomposition of a connected graph into its connected components implies

$$G(x,y) = e^{C(x,y)}, \tag{2.7}$$

and equation that has no analog for maps since maps are connected by definition. Solving (2.4), (2.5) and (2.6) explicitly is a non-trivial problem. It was done in [42] by finding an explicit expression for $B(x,y)$ in terms of $D(x,y)$; it is worth remarking that the same solution can be recovered in a more combinatorial way [19]. From this expression one can determine $C(x,y)$ as the solution of (2.6) and then $G(x,y)$ from (2.7), thus solving completely the problem of enumerating planar graphs. In particular, the estimate in (1.2) is obtained. We will not go into the details, which are quite technical, but rather will explain how the solution from [42] opened the way to the fine analysis of random planar graphs.

3. Random planar graphs

In addition to the enumerative theory of planar maps, a number of relevant results on random maps where established by Bender, Gao, Richmond and Wormald, among others. A central result in [6] is that a random map almost surely contains linearly many copies of any given planar submap M. This was later refined by showing that the number of copies of M is asymptotically normal [40]. These results extend to several classes of maps, such as triangulations and quadrangulations. Another result is that the distribution of vertex degrees follows asymptotically a discrete law with exponential tail; this already follows from Tutte's equations, and later it was shown that the limiting distribution is independent of the surface [37]. A very precise result was obtained for the distribution of the maximum vertex degree [39], proving that it is of order $\log_{6/5} n$ for maps with n edges. In another direction, it was shown [38] that a random map contains a unique 2-connected component of linear size, more precisely of size $n/3$, a result that extends to more general kind of 'components' in different classes of maps. The limiting distribution of the size of the largest component was obtained in [3], showing that it is non-Gaussian. With respect to metric properties of maps, it was first established in [21] that the typical distance between two vertices in a random quadrangulation is of order $n^{1/4}$. As we discuss later this has led to a rich theory of scaling limits of random maps. We will review several of these results when discussing random planar graphs.

The first attempt to analyze random planar *graphs* (without and embedding) was made by in [23]. The probabilistic model is given by the set $\mathcal{G}_n$ of (labelled) planar graphs with n vertices equipped with the uniform distribution. The goal declared there was to understand 'what does a random planar graph look like' under this distribution. The authors proved a

few preliminary results but it was not until the work of McDiarmid, Steger and Welsh [59] that more significant results were obtained. Among other results, they proved that a random planar graph has with high probability linearly many disjoint *pendant* copies of each fixed connected planar graph H with a distinguished vertex v: a pendant copy of H is a subgraph isomorphic to H joined to the rest of the graph through a single edge uv, and such that the isomorphism respects the order of the labels (so that automorphisms are not considered). This implies in particular that there are linearly many vertices of degree k, for each fixed $k \geq 1$. It also implies that a random planar graph has exponentially many automorphisms (consider pendant copies of $K_{1,2}$ rooted at the vertex of degree two), in sharp contrast with arbitrary random graphs. Another property proved in [59] is that the limiting probability p that a random planar graph is connected is bounded away from 0 and from 1. In particular, it was proved that $p \geq e^{-1}$. At about the same time several authors studied the number of edges in random planar graphs. Using various combinatorial arguments it was proved that almost surely the number of edges is between $1.85n$ and $2.44n$, but no concentration result or limiting distribution was obtained.

The results in [42] allow for a much more precise description. Let $G_{n,m,k}$ be the number of planar graphs with n vertices, m edges, and k components. The key fact is that it is possible to find an *exact* expression for the exponential generating function

$$G(x, y, u) = \sum_{n,m,k \geq 0} G_{n,m,k}\, y^m u^k \frac{x^n}{n!}.$$

As we have seen before, exact does not mean simple. However the series $G(x, y, u)$ can be expressed in terms of the solution of the system of equations (2.4–2.7) involving only elementary functions and the generation function $T(x, y)$ of 3-connected rooted maps counted according to vertices and edges, which is algebraic of degree four. Everything is explicit and computable with the help a computer algebra system; see [43] for a detailed survey.

Let X_n be the random variable equal to the number of edges in planar graphs with n vertices. The distribution of X_n is completely encoded in the generating function $A(x, y) = G(x, y, 1)$, since the probability generating function of X_n is simply

$$p_n(y) = \frac{[x^n]A(x, y)}{[x^n]A(x, 1)},$$

where $[x^n]$ denotes the coefficient of x^n. From the system of equations satisfied by G it is possible to extract, using analytic methods, information on the rate of growth of its coefficients. In this case one proves that, for fixed $y > 0$, we have the estimate

$$[x^n]A(x, y) \sim c(y) n^{-7/2} \gamma(y)^n n!.$$

This already gives the estimate (1.2) with $c = c(1)$ and $\gamma = \gamma(1)$, but it gives more, namely

$$p_n(y) = \frac{c(y)}{c(1)} \left(\frac{\gamma(y)}{\gamma(1)} \right)^n + O\left(\frac{1}{n} \right), \tag{3.1}$$

where the error term comes from the method of singularity analysis used in deriving the estimates. The probability generating function is close to being an exact power and extensions of the Central Limit Theorem imply a Gaussian limit law for X_n (see Section IX.5 in [35]). Moreover, from (3.1) it follows that the expected value $\mathbb{E}X_n = p_n'(1)$ is asymptotically

μn, were $\mu = (\gamma'(y)/\gamma(1))$. The function $\gamma(y)$ is analytic and computable and one obtains $\mu \approx 2.21$. A similar computation gives $\mathrm{Var}(X_n) \approx 0.43n$. It is also possible to prove a Local Limit Theorem and to show that $\mathbb{P}(|X_n - \mathbb{E}X_n| > \epsilon n)$ is exponentially small [42]. This gives a very precise picture of the distribution of the number of edges, highly concentrated around μn.

The analysis of the number of components is easier. We have

$$G(x, 1, u) = e^{uC(x)}, \tag{3.2}$$

where $C(x)$ is the generating function of connected planar graphs, and the generating function of graphs with exactly k components is $C(x)^k/k!$. One shows that

$$\frac{[x^n]\dfrac{1}{k!}C(x)^k}{[x^n]G(x)} \sim \frac{\lambda^{k-1}}{(k-1)!}\, e^{-\lambda},$$

where $\lambda = C(\gamma^{-1}) \approx 0.037$. It follows that the random variable equal to the number of components in planar graphs with n vertices converges to $1 + \mathrm{Po}(\lambda)$, a Poisson distribution of parameter λ. In particular, the limiting probability of connectedness is $p = e^{-\lambda} \approx 0.96$. As will be seen in the next section, the largest component contains almost all vertices: its expected size is $n - c$, where c is a small constant. The small number of vertices not in the largest component accounts for the fact that $p < 1$.

Another parameter of interest is the distribution of the vertex degrees. For fixed $k \geq 1$, let $X_{k,n}$ be the number of vertices of degree k in planar graphs with n vertices. As mentioned before, $X_{k,n}$ is linear in n with high probability. It is natural then to except that $\mathbb{E}X_{k,n} \sim p_k n$ as $n \to \infty$. However, much technical work is needed in order to prove this result. It requires a very fine analysis of the generating function $G^\bullet(x, w)$ of graphs with a distinguished vertex (the root), where w marks the degree of the root. It is proved in [27] that the p_k indeed exist and that $\sum_{k\geq 1} p_k = 1$. This is equivalent to saying that the probability that a random vertex in a planar graph has degree k tends to p_k as $n \to \infty$, and that the degree distribution converges to a discrete law. The explicit expression for the probability generating function is extremely involved but it is computable and one obtains the first values

$$p_1 \approx 0.037, \quad p_2 \approx 0.16, \quad p_3 \approx 0.24, \quad p_4 \approx 0.19, \quad p_5 \approx 0.13, \quad p_6 \approx 0.09.$$

The distribution decays exponentially like $p_k \sim q^k k^{-1/2}$, where $q \approx 0.67$ is an explicit constant, suggesting that the maximum degree is asymptotically $\log_{q^{-1}}(n)$. This is indeed the case as discussed in the next section.

It was proved using different methods [66] that the number of vertices of degree k is concentrated around its expected value. It is thus natural to expect that $X_{k,n}$ is asymptotically normal as $n \to \infty$, but this is still an open problem. On the other hand, asymptotic normality of the $X_{k,n}$ has been established for simpler classes of graphs [26] as well as for planar maps [30].

4. Extremal parameters

In this section we focus on several extremal parameters that have been successfully analyzed for random planar graphs. Other extremal parameters will be discussed in the last section.

Largest component. Let us start with an easy parameter, the size L_n of the largest connected component. It has already been mentioned that L_n is almost equal to n, but we can be more precise. Let G_n be the number of planar graphs and C_n the number of connected planar graphs. The probability that $L_n = n-k$, for fixed k and $n > 2k$, is $\binom{n}{k} C_{n-k} G_k / G_n$, since there are $\binom{n}{k}$ ways of choosing the labels of the vertices not in the largest component, C_{n-k} ways of choosing the largest component, and G_k ways of choosing the complement. Using the known estimates for G_n and C_n we arrive at

$$\mathbb{P}(L_n = n-k) \sim p \cdot G_k \frac{\gamma^{-k}}{k!}, \tag{4.1}$$

where p is the limiting probability of connectivity. Because of (1.2), this quantity is of order $k^{-7/2}$ for large k. It follows that $n - L_n$ has a limiting discrete distribution with constant expectation and variance. The expected value is computable and $\mathbb{E}(n - L_n) \approx 0.038$. Readers familiar with the giant component phenomenon in the $G(n, M)$ model may wonder about analogs for planar graphs; this will be discussed in the last section.

We can also find the limiting distribution of the *fragment*, the complement of the largest component. The probability that the fragment is isomorphic to a given unlabelled graph H with h vertices is, for $n > h$,

$$\binom{n}{h} \frac{h!}{\mathrm{aut}(H)} \frac{C_{n-h}}{G_n},$$

where $\mathrm{aut}(H)$ is the number of automorphisms of H and $h!/\mathrm{aut}(H)$ is the number of different labellings of H. It follows as before that

$$\mathbb{P}(\text{fragment } \cong H) \sim p \frac{\gamma^{-k}}{\mathrm{aut}(H)}. \tag{4.2}$$

We will see in Section 6 that this result holds in a more general context.

Largest block. Because of the previous result, from now on we focus on *connected* planar graphs. A connected graph decomposes into blocks, which are either single edges (isthmuses) or maximal 2-connected subgraphs. It is natural to consider the size of the largest block. This is a very interesting parameter that has a non-Gaussian continuous limit law. It was first studied for random maps in [38], where it was proved that the largest block in a random map with n edges has expected size $\sim n/3$ and, moreover, the second largest block is of order $O(n^{2/3})$. This result is somehow comparable to the classical giant component phenomenon, a random map has a unique block of linear size and the other blocks are small. The limiting distribution for the size X_n of the largest block in random maps was determined very precisely in [3], and it involves the density function $g(x)$ of a stable law of parameter $3/2$. The precise result is the following:

$$\mathbb{P}(X_n = \lfloor n/3 + xn^{2/3} \rfloor) \sim g(x) n^{-2/3}, \tag{4.3}$$

uniformly for x in any bounded interval. That is, the largest block has expected size $n/3$ and fluctuations of order $O(n^{2/3})$. It is worth remarking that the distribution has no second moment and is asymmetric: the left tail (as $x \to -\infty$) decays polynomially while the right tail (as $x \to +\infty$) decays exponentially. The proof is based on analyzing the size of the root-block, that is, the block containing the root edge. Equation (2.2) is the basis of the analysis:

the composition scheme $B(zM(z)^2)$ is *critical*, in the sense that the evaluation of $zM(z)^2$ at its singularity $1/12$ is precisely $4/27$, which is the singularity of $B(z)$. Everything boils down then to estimating the coefficients of large powers of generating functions, which is achieved by a delicate application of the saddle-point method.

Using the tools developed in [3], an analogous result was proved for random planar graphs [44]. In this case the expected size of the largest block (now n is the number of vertices) is $\sim \alpha n$, where $\alpha \approx 0.96$ (this value of α was obtained independently in [65] using alternative methods). The limiting distribution is of the same kind as (4.3), but with a different scaling of $g(x)$. The results in [44] also give the limiting distribution for the size of the largest 3-connected component in random connected planar graphs, which again is of the same kind as (4.3), both in the number of vertices and in the number of edges. The expected number of vertices in the largest 3-connected component is $\sim 0.73n$, and the expected number of edges is $\sim 1.79n$.

A parameter related to the largest block is the following. The *2-core* of a graph G is the maximum subgraph C with minimum degree at least two. The 2-core C is obtained from G by repeatedly removing vertices of degree one and, conversely, G is obtained by attaching rooted trees at the vertices of C. It is proved in [64] that the size of the 2-core of a random planar graph is asymptotically Gaussian with expectation $\sim 0.962n$ (the value of the constant was previously found in [57]). The constant is a bit larger than the value 0.96 for the largest block; this is consistent since the 2-core clearly contains the largest block. It is also proved in [64] that the size of the largest tree attached to the 2-core is of order $c \log n$ where $c \approx 0.43$.

Maximum degree. Let Δ_n be the maximum degree in random planar graphs. A simple and elegant argument by McDiarmid and Reed [58] based on double counting and elementary properties of random planar graphs shows that with high probability

$$c_1 \log n \leq \Delta_n \leq c_2 \log n,$$

for some positive constants c_1 and c_2. This already gives the right order of magnitude. Analytic methods are needed in order to obtain a more precise result. From the previous section we know that there is a limiting degree distribution $\{p_k\}_{k\geq 1}$ with tail of order $q^k k^{-1/2}$. Using the first moment method and analytic properties of the generating function $G^\bullet(x,w)$ mentioned in the previous section, one can show that $\Delta_n \leq (1+o(1))c \log n$, where $c = 1/\log(q^{-1}) \approx 2.53$. In principle a matching lower bound could be proved using the second moment method, by rooting at a secondary vertex in addition to the root vertex. This is done in [28] for simpler classes of graphs, which is already very demanding. However, the technical difficulties with this approach for planar graphs appear insurmountable, since the equations defining the associated generating functions are just too complicated.

In order to obtain a lower bound one can use *Boltzmann samplers*, introduced in [31] for the random generation of combinatorial objects. If $\mathcal{A}$ is a class of combinatorial objects with generating function $A(x)$, and x_0 is such that $A(x_0)$ is convergent, then an object $\alpha \in \mathcal{A}$ of size n is assigned probability $x_0^n/A(x_0)$. The objects generated fluctuate in size, but all the objects of size n have the same probability.

This framework has been applied successfully since then, in particular to the efficient generation of random planar graphs [36]. One can use Boltzmann samplers not only for random generation but also for the analysis of random combinatorial objects. This approach has proved useful in particular for random planar graphs [65, 66]. This is also the case here,

using the fact that there is a unique block of linear size: a typical random planar graph G can be thought of as a large block B together with small planar graphs attached to its vertices. If we later condition on the total size of G being n, we may start with the graphs attached to B being drawn independently from the set of all connected planar graphs. In this way one recovers the power of independent samples allowing to use techniques closer to the classical theory of random graphs. This program has been carried out in [29], showing that with high probability

$$|\Delta_n - c \log n| = O(\log \log n),$$

and

$$\mathbb{E}(\Delta_n) = (1 + o(1))\, c \log n.$$

Diameter. Let D_n denote the diameter of a random connected planar graph. This is a difficult parameter to analyze, even for relatively simple classes of graphs, such as trees. The starting point is the analysis of metric properties of random planar maps, by now a rich and deep theory with connections to physics and other areas. Let Q_n be a random embedded quadrangulation (all faces of degree four) with n faces and let r_n be the radius (maximum graph distance) in Q_n with respect to a fixed base point. In the pioneering work [21] it was shown that r_n is of order $n^{1/4}$, in fact, much more was proved: $r_n/n^{1/4}$ converges in law to a continuous distribution related to Brownian motion. Notice that the diameter of Q_n is between r_n and $2r_n$. The proof in [21] is based on a bijection between quadrangulations and plane trees enriched with labels that keep track of the distances in Q_n. The typical height of a tree is of order $\sqrt{n}$, and the labels behave like a random walk along the branches of the tree. This implies that the maximum distance is of order $(\sqrt{n})^{1/2}$, explaining the exponent $1/4$. These results were later extended to other classes of random maps and, more recently, even deeper results have been established. If one consider Q_n as a metric space with the graph distance d_n, then $(Q_n, d_n n^{-1/4})$ converges in a precise technical sense to a certain random compact metric space, known as the Brownian map [52, 61].

One can use the former results to analyze the diameter D_n in random planar graphs. This is done in [17] starting from the result on quadrangulations and then moving to maps with increasing connectivity. Once a result is established for 3-connected maps, it can be transferred to 3-connected planar graphs and then to connected planar graphs. One uses in an essential way the existence of a giant block and 3-connected component, both in maps and graphs. The price to pay in this scheme for transferring the results from maps to graphs is a loss in precision. The result proved in [17] is that for $\epsilon > 0$ small enough and n large enough,

$$\mathbb{P}(D_n \in (n^{1/4-\epsilon}, n^{1/4+\epsilon})) \geq 1 - \exp(-n^{c\epsilon}).$$

It is natural to conjecture that the radius r_n of connected planar graphs scaled by $n^{-1/4}$ converges to the same law as for quadrangulations and other classes of maps, but much more precise results are needed in order to prove such a statement.

Summary of results. From this and the previous section we can conclude that we have now a rather complete picture of 'what a random planar graph looks like'. We summarize the main properties in the following list. All the results are understood to hold asymptotically almost surely when $n \to \infty$. All the constants are explicit and computable to any desired precision. The values given are approximations.

1. The *number of edges* is Gaussian with expectation $2.21n$ and linear variance.
2. The *number of connected components* is $1 + \mathrm{Po}(0.037)$. The probability of being connected is 0.96.
3. If L_n denotes the *size of the largest component*, then $n - L_n$ follows a discrete law. The expected value of $n - L_n$ is 0.38.
4. For each fixed connected planar graph H rooted at a distinguished vertex, the number of pendant copies of H is Gaussian with expectation $(\gamma^{-h}/h!)n$ and linear variance.
5. The chromatic number is four. This follows from the Four Colour Theorem and the fact that it contains K_4 as a subgraph.
6. The *number of automorphisms* is exponential in n.
7. The *number of blocks* is Gaussian with expectation $0.039n$. The *number of cut vertices* is Gaussian with expectation $0.038n$. In both cases the variance is linear.
8. For each fixed 2-connected planar graph L, the *number of blocks isomorphic to* L is Gaussian with linear expectation and variance.
9. For each $k \geq 1$, the expected number of *vertices of degree* k is $p_k n$, where the p_k are computable and $\sum p_k = 1$.
10. The *maximum degree* satisfies $|\Delta_n - c \log n| = O(\log \log n)$, where $c = 2.53$, and $\mathbb{E}\Delta_n \sim c \log n$.
11. The size of the *largest block* has expected value $0.96n$ and follows a stable law of parameter $3/2$. The remaining blocks are of size $O(n^{2/3})$. The same holds for the size of the largest 3-connected component, with expectation $0.73n$.
12. The size of the *2-core* is Gaussian with expectation $0.962n$ and linear variance.
13. The *diameter* D_n is in $(n^{1/4-\epsilon}, n^{1/4+\epsilon})$ with high probability.

5. Graphs on surfaces

The theory of map enumeration extends to maps on surfaces. A map on a surface S is a 2-cell embedding (all faces must be homeomorphic to disks) of a connected graph in S. It is worth remarking that a map on an orientable surface can be encoded in a purely combinatorial way by means of a rotation system, which consists of giving a cyclic ordering of the edges around each vertex. By giving appropriate signs to the edges the encoding also works for non-orientable surfaces, but for conciseness we only discuss the orientable case [62]. Let M_n^g be the number of maps with n edges on the orientable surface of genus g. As opposed to the planar case, there is no closed formula for M_n^g, but one can use Tutte's methodology of removing the root edge to find the associated generating function $M^g(z)$. Using induction on the genus, it was proved by Bender and Canfield [4] that $M^g(z)$ is a rational function in $\sqrt{1-12z}$. The explicit expression is quite involved but it can be used to prove the estimate

$$M_n^g \sim c_g n^{5(g-1)/2} 12^n. \tag{5.1}$$

Notice that the genus only affects the subexponential term and not the exponential growth. The surprising exponent $5(g-1)/2$ was later explained more combinatorially in [20].

Suppose one wishes, as for planar graphs, to use the enumeration of maps on S for counting graphs (without an embedding) on S. There are two main obstacles for this program: 1) no degree of connectivity guarantees a unique embedding, and 2) the class of graphs embeddable in S is not close under taking connected components or blocks, so that the basic equations among generating functions, such as (3.2) no longer hold. The road to the solution, found independently in [18] and [7], is the following. The *face-width* of a map M in S is the minimum number of intersections of M with a simple non-contractible curve C on S. It is easy to see that this minimum is achieved when C meets M only at vertices. Face-width is in some sense a measure of local planarity, if the face-width is large then the embedding is locally planar in large balls. The face-width of a graph G is the maximum face-width among all the embeddings of G.

The key result is that a 3-connected graph with large enough face-width has a unique embedding [62]. It turns out that the generating series of 3-connected maps of any fixed face-width has a negligible contribution in the asymptotic analysis [8]. Therefore, the enumeration of 3-connected graphs in a surface S can be reduced, up to negligible terms, to the enumeration of 3-connected maps in S. There is one technical difficulty, which is to enumerate maps according to edges and a suitable weight on the vertices. This is achieved starting with the enumeration of all maps in S and then, using Tutte's approach based on substitution, going to 2-connected and then to 3-connected maps in S. It is important to remark that, since maps with small face-width are discarded, one does not work with the exact counting series. Instead, if $f(x)$ is the series of interest, one finds computable series $f_1(x)$ and $f_2(x)$ such that $f_1(x) \preceq f(x) \preceq f_2(x)$ (where $\preceq$ means coefficient-wise inequality) and $f_1(x)$ and $f_2(x)$ have the same leading asymptotic estimates.

For the second obstacle one can use a result from [69]: if a connected graph G of genus g has face-width at least two, then G has a unique block of genus g and the remaining blocks are planar. A similar result holds for 2-connected graphs and 3-connected components. Since for planar graphs we have exact expressions for all the generating functions involved, starting from the (asymptotic) enumeration of 3-connected graphs of genus g we can achieve the enumeration of all graphs of genus g. Let us make more precise one of the steps in the analysis. Let $G^g(x)$ and $C^g(x)$ be the generating functions of graphs and connected graphs of genus at most g, respectively. The usual relation $G^g(x) = \exp(C^g(x))$ does not hold, since the union of graphs of genus g will have larger genus if $g > 0$. Instead, we have

$$G^g(x) \sim C^g(x) e^{C^0(x)},$$

where the symbol $\sim$ must be understood as the fact that the two functions have the same dominant terms in their singular expansions. Similarly, the relation between $C^g(x)$ and the generating function $B^g(x)$ of 2-connected graphs of genus g is not an exact equation as in the planar case, since genus is also additive in blocks, but rather an approximate version. The technical details are involved but the essence is to discard maps and graphs with small face-width.

In addition, the former approach allows one to analyze parameters of a random graph embeddable in the surface S^g of genus g. All the main parameters behave as in the planar case: number of edges is Gaussian with the same moments, number of components is 1 plus a Poisson law with the same parameter, the size of the largest component follows the same law as in (4.1), and the size of the largest 2-connected and 3-connected components obey stable laws with the same expectations. In addition, a random graph embeddable in S^g almost surely does not embed in a simpler surface. Thus we have a clear picture of what a

random graph embeddable in S^g looks like. It has a unique largest component C of genus g and the remaining components are planar. Within C there is a unique block B of linear size that has genus g and the remaining blocks are planar. Finally, B has a unique linear 3-connected component T of genus g, and the remaining 3-connected components are planar. Moreover, the graph T has a unique embedding in S^g. Extremal parameters like the diameter or the maximum degree behave likely as in the planar case, but the analysis is yet to be done.

We conclude this section with a short comment. Given a connected planar graph H, a random graph in S^g contains linearly many pendant copies of H, the proof being the same as for planar graphs. But if H is non-planar then a random graph in S^g does not contain H as a subgraph almost surely, because all the balls of radius R are planar for each fixed R. Taking R larger than the diameter of H we would reach a contradiction.

6. Minor-closed classes of graphs

We recall that a class of graphs $\mathcal{G}$ is minor-closed if whenever G is in $\mathcal{G}$ and H is a minor of G, then H is also in $\mathcal{G}$. The theory of graph minors is one of the main achievements in modern combinatorics, culminating with the great theorem of Robertson and Seymour: every minor-closed class of graphs is defined in terms of a *finite* number of excluded minors; see [53] for a quick overview. The basic example is Kuratowski's theorem, which identifies K_5 and $K_{3,3}$ as the excluded minors for planar graphs. There are several important properties that have been established for proper (excluding at least one graph) minor-closed classes of graphs. To begin with they are sparse: the number of edges is at most αn for some constant α depending only on the class. This is easy to prove with $\alpha = 2^t$, where t is the size of an excluded minor, although the correct order of magnitude of α is $t\sqrt{\log t}$ [72]. Secondly, they are *small*: the number G_n of graphs in the class with n vertices is bounded as

$$G_n \leq c^n n!,$$

for some constant $c > 0$. This implies in particular that the generating function $G(x) = \sum G_n x^n/n!$ has positive radius of convergence and defines an analytic function near 0. This was first proved in [63] and then in [32] in a more general context. Additional properties are, for example, the existence of separators of size $O(\sqrt{n})$ and the fact that the tree-width is $O(\sqrt{n})$ [2].

The systematic study of *random* graphs from a minor-closed class is more recent. Let $\mathcal{G}$ be a proper minor-closed class which is *addable*. This means that 1) a graph G is in $\mathcal{G}$ if and only the connected components of G are in $\mathcal{G}$; 2) for each graph G in $\mathcal{G}$, if u and v are vertices in different components of G, the graph obtained by adding an edge joining u and v is also in $\mathcal{G}$. This is equivalent to the condition that all the excluded minors of $\mathcal{G}$ are 2-connected. Planar graphs form an addable class, but graphs embeddable in a surface other than the sphere do not, since genus is additive on disjoint unions. Addable minor-closed classes are analyzed by McDiarmid in [56]. The first property, already proved in [60], is the existence of a growth constant γ, which is the limit

$$\gamma = \lim_{n\to\infty} \left(\frac{G_n}{n!}\right)^{1/n}. \tag{6.1}$$

In fact, more is true. The class $\mathcal{G}$ is called *smooth* if

$$\lim_{n\to\infty} \frac{G_n}{nG_{n-1}} = \gamma. \tag{6.2}$$

Of course, if the former limit exists it must equal γ, but condition (6.2) is stronger than (6.1). It is shown in [56] that addable minor-closed classes are smooth. This is proved using the 2-core discussed before and applying a technique from [5].

From now on $\mathcal{G}$ is an addable minor-closed class and R_n is a random graph from $\mathcal{G}$ with n vertices under the uniform distribution. Several basic properties have been established for R_n. If was already proved in [60] that R_n contains a linear number of pendant copies of every fixed connected graph H in $\mathcal{G}$. Using the smoothness condition this is strengthened in [56], as follows. If X_n is the number of pendant copies of H in R_n, then

$$\frac{X_n}{n} \to \frac{\gamma^{-h}}{h!} \qquad \text{in probability,} \tag{6.3}$$

where h is the number of vertices in H. In the case of planar graphs a stronger result is shown in [42] using analytic methods, namely that X_n is asymptotically Gaussian with expectation $(\gamma^{-h}/h!)n$. The great interest of the less precise result (6.3) is that it holds for *every* addable minor-closed class, where generating functions are seldom available. In particular, we deduce that for each $k \geq 1$ there is a linear number of vertices of degree k, and that the number of automorphisms is exponential.

The more precise results from [56] are on the structure and number of connected components. Let $\rho = \gamma^{-1}$, which is the radius of convergence of the counting generating function $G(x)$. We have $0 < \rho \leq 1/e$. The first inequality because $\mathcal{G}$ is small, and the second one because $\mathcal{G}$ contains the class of forests, which grows exponentially like $e^n n!$. It also holds that $G(\rho)$ is finite [56]. Let now $\mathcal{C}$ be the set of connected graphs in $\mathcal{G}$, and let $C(x)$ be the associated generating function. From general enumerative principles [35] we have the relation $G(x) = \exp C(x)$, and it follows that $C(\rho)$ is finite too. Denote by L_n the size of the largest component. We can now describe the main results from [56]. As before, all statements hold asymptotically almost surely. For a given graph H, we denote the number of vertices by $|H|$.

1. The number of components is distributed like $1 + \mathrm{Po}(C(\rho))$. In particular, the probability of connectedness is $e^{-C(\rho)}$.
2. For distinct unlabelled connected graphs $H_1, \ldots, H_k$ in $\mathcal{G}$, the numbers of components X_i isomorphic to H_i are asymptotically independent with distribution $\mathrm{Po}(\lambda_i)$, with $\lambda_i = \rho^{|H_i|}/\mathrm{aut}(H_i)$.
3. $n - L_n$ follows a discrete law. For each fixed k,
$$\mathbb{P}(n - L_n = k) \to \frac{1}{G(\rho)} \frac{G_k \rho^k}{k!}.$$
4. Given a fixed graph H, the probability that the fragment (the complement of the largest component) is isomorphic to H tends to
$$\frac{\rho^{|H|}}{\mathrm{aut}(H)\, G(\rho)}.$$

Notice that item 3 corresponds exactly to equation (4.1), since $\rho = \gamma^{-1}$ and $1/G(\rho) = e^{-C(\rho)}$ is the probability of being connected. The same applies to item 4 with respect to (4.2). Which values are possible for the limiting probability of connectivity $e^{-C(\rho)}$? It was conjectured in [60] that, among all addable classes, this probability is minimized for the class of forests, in which case it is $e^{-1/2}$. This conjecture has been proved independently in [1] and [48].

For other parameters of interest, like the number of edges, there are no general results available. The number of edges is linear by the general bound on minor-closed classes, but we do not know hot to prove, for instance, any concentration result. The same goes for the number of vertices of given degree and other basic parameters. Adapting the techniques from [58], it is proved in [41] that for addable classes whose excluded minors are all 3-connected, the maximum degree Δ_n is at least $c \log n$ for some constant $c > 0$ (this does not apply, for instance, to the class of forests, where $\Delta_n \sim \log n / \log \log n$). For any addable minor-closed class it is conjectured that $\Delta_n \leq c' \log n$, but the proof of the upper bound for planar graphs in [58] does not extend to the general case.

For non-addable classes there are few general results, but some very interesting examples. Let $\mathcal{G}_k$ be the class of graphs containing at most k disjoint cycles. This class is minor-closed but not addable. Let $\mathcal{F}_k$ be the class of graphs G such that removing k vertices from G the graph becomes a forest. In other words, graphs in $\mathcal{F}_k$ are obtained from a forest F by adding k new vertices and connecting them in any way to F. Clearly $\mathcal{F}_k \subset \mathcal{G}_k$. It is proved in [50] that almost every graph in $\mathcal{G}_k$ is in $\mathcal{F}_k$, as $n \to \infty$. This gives in particular the asymptotic growth of $\mathcal{G}_k$, since it can be shown that the number of graphs in $\mathcal{F}_k$ grows like

$$c_k 2^{kn} f_n,$$

where f_n is the number of forests and c_k is an explicit constant. The simple structure of graphs in $\mathcal{F}_k$ also gives access to properties of random graphs from $\mathcal{G}_k$. This approach has been generalized to other classes excluding disjoint copies of a given family of graphs [51].

Another example of a non-addable class is the class $\mathcal{A}$ of graphs whose components are caterpillars; a caterpillar is a tree obtained from a path by adding leaves. This class and related classes can be analyzed using generating functions [16]. It is proved, for instance, that the number of components in $\mathcal{A}$ follows a Gaussian law with expectation of order $\sqrt{n}$, a very different behaviour from what we have seen in addable classes.

To conclude this section, we mention a recent result on logical limit laws [45]. Consider a graph property expressible in first order (FO) logic, for example the existence of a triangle or the existence of an isolated vertex. Given a class of graphs $\mathcal{G}$, we are interested in the limiting probability $p(\phi)$, as $n \to \infty$, that a FO formula ϕ is satisfied in $\mathcal{G}$, provided this limit exists. This problem has been much studied for the random graph $G(n, p)$. One of the earliest results is that for constant p (in particular $p = 1/2$, the uniform model on labelled graphs), for every first order property ϕ we have either $p(\phi) = 0$ or $p(\phi) = 1$. This is called a zero-one law (see [71] for much more in this area). Zero-one laws have been studied for other combinatorial structures, such as permutations or partitions [22], and also for maps on surfaces [6]. More recently, a zero-one law was proved for random labelled trees [54]. Moreover, it holds for properties expressible in the richer monadic second order (MSO) logic, in which we are allowed to quantify over sets of vertices, in addition to quantifying over vertices. Properties such as connectivity or k-colorability can be expressed in MSO but not in FO. It is proved in [45] that for every addable minor-closed class $\mathcal{G}$ and every MSO formula ϕ, the limiting probability $p(\phi)$ exists. Moreover, if we restrict to con-

nected graphs in $\mathcal{G}$, then a zero-one law holds. It is also proved that the closure of the set $\{p(\phi) \mid \phi \text{ MSO formula}\}$ of limiting probabilities is a finite union of at least two intervals in $[0, 1]$. For the class of planar graphs, the set of intervals is completely determined.

7. Subcritical classes

For the next definition we need a bit more on generating functions. Let $\mathcal{G}$ be a class of graphs which is *block-stable*, that is, a graph G is in $\mathcal{G}$ if and only if each of the blocks of G is in $\mathcal{G}$. This is the case, for instance, for addable minor-closed classes defined in the previous section. In this situation, as we saw in Section 2, the generating functions $C(x)$ and $B(x)$ of connected and 2-connected graphs in $\mathcal{G}$ satisfy

$$C^\bullet(x) = xe^{B'(C^\bullet(x))}, \tag{7.1}$$

where $C^\bullet(x) = xC'(x)$ is the generating function of connected graphs rooted at a vertex. Let ρ_C and ρ_B be, respectively, the radius of convergence of $C(x)$ and $B(x)$. We say that $\mathcal{G}$ is subcritical if

$$C^\bullet(\rho) < \rho_B.$$

This implies that the singular behaviour of $C(x)$ is dictated by the existence of a critical point when solving (7.1), and not by the singular behaviour of $B(x)$ at ρ_B. In fact, the critical point is the solution of $xB''(x) = 1$. The class of planar graphs is *critical*, since in this case $C^\bullet(\rho) = \rho_B$. It is clear that this is a delicate condition, since it depends on whether a certain evaluation of an analytic function is smaller than or equal than another value. Unless we have access to the generating functions, it seems that we cannot prove whether a given class is subcritical or not.

A basic example of a subcritical class is the class of series-parallel graphs; they can be characterized in several ways, among them as the graphs not containing K_4 as a minor. This class is subcritical, as shown first in [11]. Other examples are outerplanar graphs, acyclic graphs (forests), and cacti graphs (graphs whose blocks are cycles). As shown in [44], the class of graphs not containing H as a minor is subcritical in several other cases, including $H = K_5 - e$ (the complete graph K_5 minus an edge). A general framework was introduced in [44] for analyzing block-stable classes of graphs whose 3-connected components are predefined. A fundamental dichotomy was found (see also [65]) between critical and subcritical classes. As we have seen, a random planar graph has a block of linear size. In contrast to this, a random graph from a subcritical class has blocks of size $O(\log n)$ and the block size follows a discrete distribution. In a sense, subcritical classes are close to trees: a typical graph is made of a linear number of small blocks forming a tree whose height is of order $\sqrt{n}$. We discuss further this dichotomy in the last section.

With respect to other parameters such as the number of edges or the number of components, the behaviour is the same for critical and subcritical classes. It is worth remarking that the only examples we know of critical classes are planar graphs and classes very close to them, such as graphs not containing $\mathcal{K}_{3,3}$ as a minor [44]. A systematic study of subcritical classes was done in [25]. It is shown that the asymptotic growth is always of the form $c \cdot n^{-5/2}\gamma^n n!$ for computable constants c and γ. Remarkably, this is also proved for the corresponding *unlabelled* classes, where symmetries have to taken into account and cycle-index sums are needed; the estimate in this case is of the form $c_u n^{-5/2}\gamma_u^n$. In addition, the number

of edges and other linear parameters are asymptotically Gaussian with linear expectation and variance.

8. Concluding remarks

In this section we discuss additional aspects of random planar graphs and related classes, and several open problems. So far we have discussed random planar graphs according to the number of vertices, but it is also interesting to consider the number $G_{n,m}$ of planar graphs with n vertices and m edges. There are two different situations. First, when $m = \alpha n$ for $\alpha \in (1, 3)$. This was addressed in [42], and it was shown that

$$G_{n,\lfloor \alpha n \rfloor} \sim c(\alpha) n^{-4} \gamma(\alpha)^n n!,$$

where $c(\alpha)$ and $\gamma(\alpha)$ are analytic functions of α. The function $\gamma(\alpha)$ has a strict maximum at $\mu \approx 2.21$, where μn is the expected number of edges. This proves in particular the large deviations result for the number of edges. It turns out that the typical behaviour of random graphs with αn edges is qualitatively the same for each $\alpha \in (1, 3)$, that is, there is no critical value of α. The matter changes if one considers $m \leq n$. As shown in [47], there are two critical periods in the 'evolution' of planar graphs with n vertices and m vertices. The first one is analogous to the phase transition observed in the standard $G(n, M)$ model and takes place for $M = n/2 + O(n^{2/3})$, when the largest complex component is formed. A second critical period appears at $n + O(n^{3/5})$, when the complex components cover nearly all vertices.

So far we have worked with labelled graphs, but all our problems make sense for unlabelled graphs as well. Let U_n be the number of unlabelled planar graphs with n vertices. We do not have yet a precise estimate for U_n, we do not even know the unlabelled growth constant $\gamma_u = \lim(U_n)^{1/n}$. Since the number of automorphisms of a random labelled planar graph is exponential, we must have $\gamma_u > \gamma = 27.23$. On the other hand, the best upper bound available is $\gamma_u < 30.06$, proved in [15]. Using Pólya's theory of counting, unlabelled graphs can be enumerated for subcritical classes [25]. In principle this could be doable for planar graphs starting at 3-connected planar graphs, but the analysis of symmetries appears too involved. In any case, one should expect an asymptotic estimate of the form $U_n \sim cn^{-7/2}(\gamma_u)^n$. Also, random unlabelled planar graphs should share the same properties as their labelled counterpart.

Another open problem we address is the possible dichotomy discussed in the previous section between critical and subcritical classes. Let $\mathrm{Ex}(H_1, \ldots, H_k)$ be the class of graphs not containing any of the H_i as a minor. For instance, $\mathrm{Ex}(K_5, K_{3,3})$ is the class of planar graphs and $\mathrm{Ex}(K_4)$ is the class of series-parallel graphs. In all cases where analytic methods are available, one observes that the class is subcritical if and only if at least one of the excluded minors is planar. A central result in the graph minors program [68], says that the tree-width of graphs in $\mathrm{Ex}(H_1, \ldots, H_k)$ is bounded if and only if at least one of the H_i is planar. The tree-width is a measure of how close is a graph to being tree-like. If we recall that graphs from subcritical classes are typically tree-like, the following conjecture seems reasonable, restricted to addable classes, where the basic equation (7.1) holds.

Conjecture. *The class* $\mathrm{Ex}(H_1, \ldots, H_k)$ *is subcritical if and only if at least one of the* H_i *is planar, which is equivalent to having bounded tree-width.*

In particular, it would be very interesting to prove this conjecture for the class $\mathcal{G}_k$ of graphs with tree-width at most k. $\mathcal{G}_1$ is the class of forests and $\mathcal{G}_2$ is the class of series-parallel graphs, which are subcritical. But already for $\mathcal{G}_3$ we do not know. We know which are the edge-maximal graphs in $\mathcal{G}_k$, the so-called k-trees. They certainly have a tree-like structure (almost by definition) but it is not clear how to infer results for random graphs in $\mathcal{G}_k$ from the maximal ones.

Another topic for future research is to analyze additional extremal parameters. The following questions refer to almost sure properties of random planar graphs.

- *Cores.* The k-core of a graph is the maximum subgraph with minimum degree at leat k. We have already discussed the 2-core, which is of linear size for random planar graphs. The 3-core is not necessarily connected, but it is conjectured [64] that the 3-core contains a component of linear size, and that the components of the 4-core are all sublinear.
- *Tree-width.* It is known that a planar graph with diameter D has tree-width $O(D)$. It follows that the tree-width is at most $O(n^{1/4+\epsilon})$. Is this the right order of magnitude? We remark that there are planar graphs (grids) with tree-width $\sqrt{n}$.
- *Longest cycle.* We conjecture the existence of cycle of length cn for some $c > 0$. Because of the results on the largest 3-connected component, it would be enough to prove it for random 3-connected planar graphs. In contrast, it is easy to see that there is always a matching of linear size (consider pendant copies of a single edge).

On the enumerative side, we mention the problem of counting 4-regular planar graphs. Cubic planar graphs can be enumerated adapting Tutte's decomposition into 3-connected components [12], but this approach does not seem to work for higher degree. In the same way, planar graphs with minimum degree three can be enumerated [64], but the same obstacle appears for minimum degree four. Another problem is to enumerate bipartite planar graphs. The real difficulty is to keep control of the bipartite character in the decomposition of 2-connected graphs into 3-connected components.

Concerning minor-closed classes of graphs, a main open problem is to show that the growth constant always exists (as conjectured in [10]). More of a metaproblem is to analyze additive parameters like the number of edges or extremal parameters like the size of the largest block in general minor-closed classes. It is not at all clear that there is a way of attacking them without precise enumerative results. One case particularly appealing is the class $\mathrm{Ex}(K_5)$. Wagner's theorem tells us how is the structure of graphs in $\mathrm{Ex}(K_5)$, but so far we are not able to obtain precise enumerative information from it.

Acknowledgements. The author is partially supported by Grants MTM2011-24097 and DGR2009-SGR1040. The author wishes to thank Michael Drmota and Colin McDiarmid for helpful comments.

References

[1] L. Addario-Berry, C. McDiarmid, and B. Reed, *Connectivity for bridge-addable monotone graph classes*, Combin. Probab. Comput. **21** (2012), 803–815.

[2] N. Alon, P. Seymour, and R. Thomas, *A separator theorem for nonplanar graphs*, J. Amer. Math. Soc. **3** (1990), 801–808.

[3] C. Banderier, P. Flajolet, G. Schaeffer, and M. Soria, *Random Maps, Coalescing Saddles, Singularity Analysis, and Airy Phenomena*, Random Strucures Algorithms **19** (2001), 194–246.

[4] E. A. Bender and E. R. Canfield, *The asymptotic number of rooted maps on a surface*, J. Combin. Theory Ser. A **43** (1986), 244–257.

[5] E. A. Bender, E. R. Canfield, and L. B. Richmond, *Coefficients of functional compositions often grow smoothly*, Electron. J. Combin. **15** #21.

[6] E. A. Bender, Z. C. Gao, and L. B. Richmond, *Submaps of maps I: General 0-1 laws*, J. Combin. Theory Ser. B **55** (1992), 104–117.

[7] E. A. Bender, Z. Gao, *Asymptotic enumeration of labelled graphs by genus*, Electron. J. Combin. **18**(1) (2011), #13.

[8] E. A. Bender, Z. Gao, L. B. Richmond, and N. Wormald, *Asymptotic properties of rooted 3-connected maps on surfaces*, J. Austral. Math. Soc., Series A **60** (1996), 31–41.

[9] E. A. Bender, Z. Gao, and N. C. Wormald, *The number of 2-connected labelled planar graphs*, Electron. J. Combin. **9** (2002), #43.

[10] O. Bernardi, M. Noy, and D. Welsh, *Growth constants of minor-closed classes of graphs*, J. Combin. Theory Ser. B **100** (2010), 468–484.

[11] M. Bodirsky, O. Giménez, M. Kang, and M. Noy, *Enumeration and limit laws for series-parallel graphs*, European J. Combin. **28** (2007), 2091–2105.

[12] M. Bodirsky, M. Kang, M. Löffler, C. McDiarmid, *Random cubic planar graphs*, Random Structures Algorithms **30** (2007), 78–94.

[13] B. Bollobás, *Random Graphs*, Academic Press, London, 1985.

[14] ———, *Hereditary and monotone properties of combinatorial structures*, Surveys in combinatorics 2007, Cambridge Univ. Press, Cambridge, 2007, pp. 1–39.

[15] N. Bonichon, C. Gavoille, N. Hanusse, D. Poulalhon, and G. Schaeffer, *Planar Graphs, via Well-Orderly Maps and Trees*, Graphs Combin. **22** (2006), 185–202.

[16] M. Bousquet-Mélou, K. Weller, *Asymptotic properties of some minor-closed classes of graphs*, Combin. Probab. Comput. (to appear).

[17] G. Chapuy, E. Fusy, O. Giménez, and M. Noy, *The diameter of random planar graphs*, Combin. Probab. Comput. (to appear).

[18] G. Chapuy, E. Fusy, O. Giménez, B. Mohar, and M. Noy, *Asymptotic enumeration and limit laws for graphs of fixed genus*, J. Combin. Theory Ser. A **118** (2011), 748–777.

[19] G. Chapuy, É. Fusy, M. Kang, and B. Shoilekova, *A complete grammar for decomposing a family of graphs into 3-connected components*, Electron. J. Combin. **15**(1) (2008), #148.

[20] G. Chapuy, M. Marcus, and G. Schaeffer, *A bijection for rooted maps on orientable surfaces*, SIAM J. Discrete Math. **23** (2009), 1587–1611.

[21] P. Chassaing and G. Schaeffer, *Random planar lattices and integrated superBrownian excursion*, Probab. Theory Relat. Fields **128** (2004), 161–212.

[22] K. J. Compton, *A logical approach to asymptotic combinatorics. I. First order properties*, Adv. Math. **65** (1987), 65–96.

[23] A. Denise, M. Vasconcellos, and D. J. A. Welsh, *The random planar graph*, Congr. Numer. **113** (1996), 61–79.

[24] M. Drmota, *Random trees*. Springer, 2009.

[25] M. Drmota, É. Fusy, M. Kang, V. Kraus, and J. Rué, *Asymptotic Study of Subcritical Graph Classes*, SIAM J. Discrete Math. **25** (2011), 1615–1651.

[26] M. Drmota, O. Giménez, and M. Noy, *Vertices of given degree in series-parallel graphs*, Random Structures Algorithms **36** (2010), 273–314.

[27] ———, *Degree distribution in random planar graphs*, J. Combin. Theory Ser. A **118** (2011), 2102–2130.

[28] ———, *The maximum degree of series-parallel graphs*, Combin. Probab. Comput. **20** (2011), 529–570.

[29] M. Drmota, O. Giménez, M. Noy, K. Panagiotou, and A. Steger, *The maximum degree of planar graphs*, Proc. London Math. Soc. (to appear).

[30] M. Drmota and K. Panagiotou, *A central limit theorem for the number of degree-k vertices in random maps*, Algorithmica **66** (2013), 741–761.

[31] P. Duchon, P. Flajolet, G. Louchard, and G. Schaeffer, *Boltzmann samplers for the random generation of combinatorial structures*, Combin. Probab. Comput. **13** (2004), 577–625.

[32] Z. Dvořák and S. Norine, *Small graph classes and bounded expansion*, J. Combin. Theory Ser. B **100** (2010), 171–175.

[33] P. Erdős, D. J. Kleitman, and B. L. Rothschild, *Asymptotic enumeration of K_n-free graphs*, International Colloquium on Combinatorial Theory, Atti dei Convegni Lincei 17, Vol. 2, Rome, 1976, pp. 19–27.

[34] P. Erdős and A. Rényi, *On the evolution of random graphs*, Publ. Math. Inst. Hung. Acad. Sci. **5** (1960) 17–61.

[35] P. Flajolet, R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2009.

[36] E. Fusy, *Uniform random sampling of planar graphs in linear time*, Random Structures Algorithms **35** (2009), 464–522.

[37] Z. Gao and L. B. Richmond, *Root vertex valency distributions of rooted maps and rooted triangulations*, European J. Combin. **15** (1994), 483–490.

[38] Z. Gao and N. C. Wormald, *The size of the largest components in random planar maps*, SIAM J. Discrete Math. **12** (1999), 217–228.

[39] _____, *The distribution of the maximum vertex degree in random planar maps*, J. Combin. Theory Ser. A **89** (2000), 201–230.

[40] Z. Gao and N. C. Wormald, *Asymptotic normality determined by high moments, and submap counts of random maps*, Probab. Theory Relat. Fields **130** (2004), 368–376.

[41] O. Giménez, D. Mitsche, and M. Noy, *Maximum degree in minor-closed classes of graphs*, (submitted), arXiv:1304.5049.

[42] O. Giménez and M. Noy, *Asymptotic enumeration and limit laws of planar graphs*, J. Amer. Math. Soc. **22** (2009), 309–329.

[43] _____, *Counting planar graphs and related families of graphs*, In Surveys in combinatorics 2009, 169–210, Cambridge Univ. Press, Cambridge, 2009.

[44] O. Giménez, M. Noy, and J. Rué, *Graph classes with given 3-connected components: asymptotic enumeration and random graphs*, Random Structures Algorithms **42** (2013), 438–479.

[45] P. Heinig, T. Müller, M. Noy, and A. Taraz, *Logical limit laws for minor-closed classes of graphs*, (submitted), arXiv:1401.7021.

[46] S. Janson, T. Łuczak, and A. Rucinski, *Random Graphs*, John Wiley, New York, 2000.

[47] M. Kang and T. Łuczak, *Two critical periods in the evolution of random planar graphs*, Trans. Amer. Math. Soc. **364** (2012), 4239–4265.

[48] M. Kang and K. Panagiotou, *On the connectivity of random graphs from addable classes*, J. Combin. Theory Ser. B **103** (2013), 306–312.

[49] P. G. Kolaitis, H. J. Prömel, and B. L. Rothschild, *K_{l+1}-free graphs: asymptotic structure and a 0-1 law*, Trans. Amer. Math. Soc. **303** (1987), pp. 637–671.

[50] V. Kurauskas and C. McDiarmid, *Random graphs with few disjoint cycles*, Combin. Probab. Comput. **20** (2011), 763–775.

[51] _____, *Random graphs containing few disjoint excluded minors*, Random Structures Algorithms **44** (2014), 240–268.

[52] J.-F. Le Gall, *Uniqueness and universality of the Brownian map*, Ann. Probab. **41** (2013), 2880–2960.

[53] L. Lovász, *Graph minor theory*, Bull. Amer. Math. Soc. **43** (2006), 75–86.

[54] G. L. McColm, *MSO zero-one laws on random labelled acyclic graphs*, Discrete Math. **254** (2002), 331–347.

[55] C. McDiarmid, *Random graphs on surfaces*, J. Combin. Theory Ser. B **98** (2008), 778–797.

[56] ———, *Random graphs from a minor-closed class*, Combin. Probab. Comput. **18** (2009), 583–599.

[57] ———, *Random Graphs from a Weighted Minor-Closed Class*, Electron. J. Combin. **20**(2) (2013), P52.

[58] C. McDiarmid and B. Reed, *On the maximum degree of a random planar graph*, Combin. Probab. Comput. **17** (2008) 591– 601.

[59] C. McDiarmid, A. Steger, and D. J. A. Welsh, *Random planar graphs*, J. Combin. Theory Ser. B **93** (2005), 187–205.

[60] ———, *Random graphs from planar and other addable classes*, Topics in discrete mathematics, Algorithms Combin., 26, Springer, Berlin, 2006, pp. 231–246.

[61] G. Miermont, *The Brownian map is the scaling limit of uniform random plane quadrangulations*, Acta Math. **210** (2013), 319–401.

[62] B. Mohar and C. Thomassen. *Graphs on surfaces*, Johns Hopkins University Press, Baltimore, MD, 2001.

[63] S. Norine, P. Seymour, R. Thomas, and P. Wollan, *Proper minor-closed families are small*, J. Combin. Theory Ser. B **96** (2006), 754–757.

[64] M. Noy and L. Ramos, *Random planar graphs with given minimum degree*, (submitted), arXiv:1403.5211.

[65] K. Panagiotou and A. Steger, *Maximal biconnected subgraphs of random planar graphs*, ACM Trans. Algorithms **6** (2010), Art. 31, 21 pp.

[66] ———, *On the Degree Distribution of Random Planar Graphs*, In: Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA '11), pp. 1198–1210.

[67] H. J. Prömel and A. Steger, *The asymptotic number of graphs not containing a fixed color-critical subgraph*, Combinatorica **12** (1992), 463–473.

[68] N. Robertson and P. D. Seymour, *Graph minors. V. Excluding a planar graph*, J. Combin. Theory Ser. B **41** (1986), 92–114.

[69] N. Robertson and R. P. Vitray, *Representativity of surface embedding*, In Paths, Flows, and VLSI-layout, Eds: B. Korte, L. Lovász, H. J. Prömel, A. Schrijver, Springer, Berlin, 1990, pp. 293–328.

[70] G. Schaeffer, *Conjugaison d'arbres et cartes combinatoires aléatoires*, Ph.D. Thesis, Université Bordeaux I, 1998.

[71] J. Spencer, *The strange logic of random graphs*, Algorithms and Combinatorics, 22. Springer-Verlag, Berlin, 2001.

[72] A. Thomason, *Extremal functions for graph minors*, More sets, graphs and numbers, Bolyai Soc. Math. Stud., 15, Springer, Berlin, 2006, pp. 359–380.

[73] W. T. Tutte, *A census of planar maps*, Canad. J. Math. **15** (1963), 249–271.

[74] N. C. Wormald, *Models of random regular graphs*, in Surveys in Combinatorics, 1999, Cambridge University Press, Cambridge (1999), pp. 239–298.

Departament de Matemàtica Aplicada. Universitat Politècnica de Catalunya., Jordi Girona 1–3, 08034 Barcelona, Spain.

E-mail: marc.noy@upc.edu

The Gelfand-Tsetlin graph and Markov processes

Grigori Olshanski

Abstract. The goal of the paper is to describe new connections between representation theory and algebraic combinatorics on one side, and probability theory on the other side.

The central result is a construction, by essentially algebraic tools, of a family of Markov processes. The common state space of these processes is an infinite dimensional (but locally compact) space Ω. It arises in representation theory as the space of indecomposable characters of the infinite-dimensional unitary group $U(\infty)$.

Alternatively, Ω can be defined in combinatorial terms as the boundary of the Gelfand-Tsetlin graph - an infinite graded graph that encodes the classical branching rule for characters of the compact unitary groups $U(N)$.

We also discuss two other topics concerning the Gelfand-Tsetlin graph:

(1) Computation of the number of trapezoidal Gelfand-Tsetlin schemes (one could also say, the number of integral points in a truncated Gelfand-Tsetlin polytope). The formula we obtain is well suited for asymptotic analysis.

(2) A degeneration procedure relating the Gelfand-Tsetlin graph to the Young graph by means of a new combinatorial object, the Young bouquet.

At the end we discuss a few related works and further developments.

Mathematics Subject Classification (2010). 05E05, 05E10, 60J27, 60J35.

Keywords. Asymptotic representation theory, representation ring, symmetric functions, Gelfand-Tsetlin graph, Young graph, Feller Markov processes, infinitesimal generators.

1. Introduction

The present paper is devoted to combinatorial and probabilistic aspects of the asymptotic representation theory. The adjective "asymptotic" means that we are interested in the limiting behavior of representations of growing groups

$$G(1) \subset G(2) \subset G(3) \subset \dots$$

and their relationship with representations of the limiting group $G(\infty)$, which is defined as the union of $G(N)$'s. Here there is a remarkable analogy with limit transitions in models of statistical physics and random matrix theory.

The model examples of the "big groups" $G(\infty)$ are the infinite symmetric group $S(\infty)$ and the infinite-dimensional unitary group $U(\infty)$. There is a striking parallelism between

Proceedings of the International Congress of Mathematicians, Seoul, 2014

the theories for these two groups that we substantially exploit. In our exposition, we focus on the more difficult case of $U(\infty)$ and only briefly mention the parallel results concerning $S(\infty)$.

The main references are the recent joint papers [11–13] by Alexei Borodin and myself, and my paper Olshanski [47]. These works originated from our previous study of the problem of harmonic analysis on $S(\infty)$ and $U(\infty)$ (Borodin-Olshanski [8–10]).

1.1. Relative dimension in the Gelfand-Tsetlin graph. All our considerations are intimately connected with the Gelfand-Tsetlin graph. We recall its definition in Section 3. As was already mentioned in the abstract, the Gelfand-Tsetlin graph encodes the branching of irreducible characters for the compact groups

$$U(1) \subset U(2) \subset U(3) \subset \dots . \tag{1.1}$$

A fundamental result in the asymptotic representation theory is the Edrei-Voiculescu theorem on the classification of indecomposable characters of the group $U(\infty)$ (Edrei [19], Voiculescu [54]). In combinatorial terms, the Edrei-Voiculescu theorem describes the *boundary* of the Gelfand-Tsetlin graph (see Section 3 for the precise definitions).

In Borodin-Olshanski [11] we propose a novel approach to this old theorem, based on the study of the *relative dimension*

$$F_{\varkappa}(\nu) := \frac{\operatorname{Dim}_{K,N}(\varkappa, \nu)}{\operatorname{Dim}_N \nu}. \tag{1.2}$$

Here $\varkappa$ and ν are two vertices of the Gelfand-Tsetlin graph, on levels K and N, respectively ($K < N$); the numerator is the number of monotone paths in the graph connecting $\varkappa$ to ν; finally, the denominator is the number of all monotone paths ending at ν (this is the same as the dimension of the irreducible character of $U(N)$ corresponding to ν). The notation in (1.2) emphasizes that we regard the ratio on the right-hand side as a function in variable ν with $\varkappa$ being a fixed parameter.

We show that $F_{\varkappa}(\nu)$ is a rather "regular" function, which shares some properties of the classic Schur functions like the examples in Macdonald [35]. What is especially important for our purposes, we obtain a good contour integral representation for $F_{\varkappa}(\nu)$, which makes it possible to find its asymptotics as ν goes to infinity.

These results are reviewed in Section 3.

1.2. The zw-measures and related Markov processes. One of the most beautiful hypergeometric identities is classic Dougall's formula (1907) which can be written as

$$\begin{aligned} &\sum_{n\in\mathbb{Z}} \frac{1}{\Gamma(z-n+1)\Gamma(z'-n+1)\Gamma(w+n+1)\Gamma(w'+n+1)} \\ &= \frac{\Gamma(z+w+z'+w'+1)}{\Gamma(z+w+1)\Gamma(z+w'+1)\Gamma(z'+w+1)\Gamma(z'+w'+1)}, \end{aligned} \tag{1.3}$$

see Dougall [17], Erdelyi [20]. Here z, z', w, w' are complex parameters such that $\Re(z+z'+w+w') > -1$ and $\Gamma(\,\cdot\,)$ is Euler's Γ-function. Let $M_{z,z',w,w'}(n)$ denote the nth summand on the left-hand side divided by the quantity on the right-hand side. It is easy to find conditions under which all the summands on the left-hand side are real and positive.

Then the quantities $M_{z,z',w,w'}(n)$, $n \in \mathbb{Z}$, define a probability measure $M_{z,z',w,w'}$ on $\mathbb{Z}$. We call it the *zw-measure*.

The zw-measures also arise in a probabilistic context. Recall that a *birth-death process* is a continuous time Markov chain (or random walk) on $\mathbb{Z}_+$ such that the only possible transitions from a state $n \in \mathbb{Z}_+$, in an infinitesimal time interval $(t, t+dt)$, are the neighboring states $n \pm 1$. Such a process is determined by specifying the jump rates $a_\pm(n)$; then the infinitesimal generator of the process is the difference operator $D_{z,z',w,w'}$ on $\mathbb{Z}_+$ acting on a test function f by

$$(Df)(n) = a_+(n)[f(n+1) - f(n)] + a_-(n)[f(n-1) - f(n)], \qquad n \in \mathbb{Z}_+.$$

The birth-death processes are well-studied objects which have many applications.

Let us now ask what could be the simplest *bilateral* analog of birth-death processes, living on the whole lattice $\mathbb{Z}$ and possessing a stationary distribution (in other words, an invariant probability measure). The generator of a bilateral process still has the same form, only the jump rates $a_\pm(n)$ are required to be strictly positive for all $n \in \mathbb{Z}$.

If the quantities $a_\pm(n)$ are constants, then there is no finite invariant measure. If $a_\pm(n)$ depends on n linearly, it changes the sign, which is inadmissible. But if we require $a_\pm(n)$ to be quadratic functions of n, then the processes with desired properties exist, they depend on four parameters, and the corresponding invariant measures are just the zw-measures.

Thus the zw-measures appear as the stationary distributions of certain natural Markov processes on $\mathbb{Z}$. Each of them is uniquely determined by its generator, which is the difference operator

$$\begin{aligned}(D_{z,z',w,w'}f)(n) = (z-n)(z'-n)(f(n+1) - f(n)) \\ + (w+n)(w'+n)(f(n-1) - f(n)). \quad (1.4)\end{aligned}$$

Here n ranges over $\mathbb{Z}$ and the parameters are subject to constraints stated in the beginning of Section 4 below.

Observe now that $\mathbb{Z}$ is the Pontryagin dual group to the unit circle

$$\mathbb{T} := \{u \in \mathbb{C} : |u| = 1\},$$

which is a commutative group isomorphic to $U(1)$. In Sections 4–6 we explain how to construct analogs of the zw-measures, the related Markov processes, and the generators $D_{z,z',w,w'}$ when $\mathbb{Z}$ is replaced by the *dual objects* to noncommutative groups $U(N)$ ($N = 2, 3, \dots$) and (which is our final goal) by the dual object to the group $U(\infty)$.

For $N = 2, 3, \dots$, the dual object $\widehat{U(N)}$, like $\widehat{U(1)} = \mathbb{Z}$, is a countable set; it is naturally identified with a subset $\mathbb{S}_N \subset \mathbb{Z}^N$ (the highest weights of the group $U(N)$). The dual object $\widehat{U(\infty)}$, on the contrary, is a continuous infinite-dimensional space: its points depend on infinitely many continuous parameters. Thus our construction leads to a four-parameter family of Markov processes on this infinite-dimensional space.

The generators of these processes are explicitly computed: they are implemented by certain infinite-variate second order partial differential operators $\mathbb{D}_{z,z',w,w'}$ (see Section 6 below). Initially, $\mathbb{D}_{z,z',w,w'}$ is defined on a certain algebra R^U — the *representation ring* of the family $\{U(N); N = 1, 2, 3 \dots\}$. As is well known, the representation ring for the family of symmetric groups is isomorphic to Sym, the algebra of symmetric functions. We regard the algebra R^U as a reasonable substitute of the algebra Sym even though R^U seems to be more sophisticated as compared to Sym.

1.3. The Young bouquet. There exists a great similarity between the representation theories of the two basic big groups, $U(\infty)$ and $S(\infty)$. It is striking when comparing the description of the dual objects (cf. Voiculescu [54] and Thoma [51]) or the construction of the generalized regular representations which are the subject of harmonic analysis (cf. Olshanski [42] and Kerov-Olshanski-Vershik [27], [28]). The study of the relative dimension (1.2) in the Gelfand-Tsetlin graph has been inspired by earlier results (Okounkov-Olshanski [39]) on the relative dimension in the Young graph — the counterpart of the Gelfand-Tsetlin graph in the symmetric group case. The zw-measures and related Markov processes also have counterparts in the symmetric group case (Borodin-Olshanski [14]).

This parallelism is in sharp contrast to the fact that the groups $U(\infty)$ and $S(\infty)$, as well as $U(N)$ and $S(N)$, look quite differently. In Borodin-Olshanski [13] we suggest an explanation of this phenomenon. The idea is that one can establish a connection between the Gelfand-Tsetlin and Young graphs by making use of a certain poset with *continuous grading*. We call this poset the *Young bouquet*.

By the very definition, the Young bouquet is a close relative of the Young graph. On the other hand, we show that the Young bouquet can be obtained from the Gelfand-Tsetlin graph by a limit transition turning the discrete grading into a continuous one. Moreover, the limit transition leads to a reasonable degeneration of various objects that are structurally connected with the Gelfand-Tsetlin graph.

We discuss the Young bouquet in Section 7. Note that the results of [13] are substantially used in the construction of Markov processes in the symmetric group case (Borodin-Olshanski [14]).

2. Dual objects and stochastic links

According to the conventional definition, the *dual object* $\widehat{G}$ to a (topological) group G is the set of equivalence classes of irreducible unitary representations of G.

For a finite or compact group, all irreducible representations have finite dimension and the dual object can be identified with the set of irreducible characters.

Let G be a compact group. For $\pi \in \widehat{G}$, denote the dimension of π by $\operatorname{Dim}\pi$. Given a closed subgroup $H \subset K$ and $\rho \in \widehat{H}$, denote by $\operatorname{Dim}(\rho,\pi)$ the multiplicity of ρ in the decomposition of $\pi\big|_H$. Counting the dimensions we get the identity

$$\operatorname{Dim}\pi = \sum_{\rho\in\widehat{H}} \operatorname{Dim}\rho \operatorname{Dim}(\rho,\pi).$$

Let us form the matrix Λ_H^G whose rows are indexed by elements $\pi \in \widehat{G}$, the columns are indexed by the elements $\rho \in \widehat{H}$, and the entries are given by

$$\Lambda_H^G(\pi,\rho) = \frac{\operatorname{Dim}\rho \operatorname{Dim}(\rho,\pi)}{\operatorname{Dim}\pi}.$$

In other words, $\Lambda_H^G(\pi,\rho)$ is the relative dimension of the ρ-isotypic component in the decomposition of $\pi\big|_H$.

Evidently, the matrix entries are nonnegative numbers, and (by virtue of the above identity) all row sums are equal to 1, so that Λ_H^G is a *stochastic matrix*. We call it a *stochastic link*

and denote by the dashed arrow, $\Lambda_H^G : \widehat{G} \dashrightarrow \widehat{H}$. Informally, we regard Λ_H^G as a "generalized map" from $\widehat{G}$ to $\widehat{H}$, dual to the inclusion map $H \to G$.

Let us return to the unitary groups $U(N)$, which are the model example of compact Lie groups. As is well known, the irreducible characters of $U(N)$, viewed as symmetric functions in the matrix eigenvalues $u_1, \dots, u_N$, have the form

$$\chi_\nu(u_1, \dots, u_N) = \frac{\det\left[u_i^{\nu_j+N-j}\right]_{i,j=1}^N}{\prod_{i,j=1}^N (u_i - u_j)},$$

where the subscript ν is an N-tuple of integers $\nu_1 \geq \dots \geq \nu_N$ called a *signature* of length N (Weyl [55], Zhelobenko [57]). Thus, the dual object $\widehat{U(N)}$ is in one-to-one correspondence with the set $\mathbb{S}_N \subset \mathbb{Z}^N$ formed by the signatures of length N.

Two signatures $\nu \in \mathbb{S}_N$ and $\lambda \in \mathbb{S}_{N-1}$ are said to be *interlaced* if their coordinates satisfy the inequalities $\nu_i \geq \lambda_i \geq \nu_{i+1}$ for every $i = 1, \dots, N-1$; then we write $\lambda \prec \nu$.

Let π_ν denote the irreducible representation with character χ_ν. The classic *Gelfand-Tsetlin branching rule* (Gelfand-Tsetlin [22], Zhelobenko [57]) says that

$$\pi_\nu\big|_{U(N-1)} = \bigoplus_{\lambda:\, \lambda \prec \nu} \pi_\lambda,$$

which is equivalent to the character relation

$$\chi_\nu(u_1, \dots, u_{N-1}, 1) = \sum_{\lambda:\, \lambda \prec \nu} \chi_\lambda(u_1, \dots, u_{N-1}).$$

Recall that the dimension of π_ν, which we denote by $\operatorname{Dim}_N \nu$, is given by the well-known *Weyl's dimension formula*

$$\operatorname{Dim}_N \nu = \prod_{1 \leq i < j \leq N} \frac{\nu_i - \nu_j - i + j}{j - i}.$$

It follows that the stochastic link $\widehat{U(N)} \dashrightarrow \widehat{U(N-1)}$ has the following form

$$\Lambda_{N-1}^N(\nu, \lambda) = \begin{cases} \dfrac{\operatorname{Dim}_{N-1} \lambda}{\operatorname{Dim}_N \nu}, & \text{if } \lambda \prec \nu, \\ 0, & \text{otherwise.} \end{cases}.$$

We explain now how we understand the dual object to the group $U(\infty)$. This group is *wild* (= not type I, see Kirillov [29, Section 8.4]), so the conventional definition of the dual object is inappropriate as it leads to a huge pathological space. For the purpose of the present work it is reasonable to adopt the following definition, which can be formulated in a few different but equivalent ways. Namely, the dual object $\widehat{U(\infty)}$ is:

Version 1. The set of quasi-equivalence classes of finite factor representations of $U(\infty)$.

This formulation follows the approach of Thoma [51] and Voiculescu [54]. Finite factor representations are uniquely (within quasi-equivalence) determined by their normalized traces, which can be characterized as indecomposable positive definite class functions $\chi : U(\infty) \to \mathbb{C}$ normalized by $\chi(e) = 1$.

Version 2. The set of functions $\chi : U(\infty) \to \mathbb{C}$ which can be approximated by the *normalized irreducible characters*

$$\widetilde{\chi}_\nu := \frac{\chi_\nu}{\chi_\nu(e)} = \frac{\chi_\nu}{\operatorname{Dim}_N \nu},$$

where we assume that $\nu \in \mathbb{S}_N$ varies together with N as N goes to infinity.

The idea of this approach is due to Vershik and Kerov [52], [53]. For more detail, see Okounkov-Olshanski [40].

Version 3. The set of positive definite class functions $\chi : U(\infty) \to \mathbb{C}$ such that $\chi(e) = 1$ and for arbitrary $g, h \in U(\infty)$ one has

$$\lim_{N\to\infty} \int_{k\in U(N)} \chi(gkhk^{-1})dk = \chi(g)\chi(h),$$

where integration is taken with respect to the normalized Haar measure on $U(N)$.

For more detail, see Olshanski [41]. The above limit relation is an analog of the classic *functional equation* for the normalized irreducible characters of compact groups.

Version 4. The categorical projective limit of the sequence of stochastic links

$$\widehat{U(1)} \leftarrow\!\!- \widehat{U(2)} \leftarrow\!\!- \widehat{U(3)} \leftarrow\!\!- \dots$$

For more detail, see Borodin-Olshanski [12], [13], Olshanski [46].

As seen from the third version above, $\widehat{U(\infty)}$ can be identified with a set of positive definite class functions on $U(\infty)$. These functions are called the *indecomposable* or *extreme characters* of $U(\infty)$. Here is their precise description.

First, notice that every element of the group $U(\infty)$ is represented by an infinite unitary matrix $g = [g_{ij}]_{i,j=1}^\infty$ such that $g_{ij} = \delta_{ij}$ when i or j is large enough. Write $u_1, u_2, \dots$ for the eigenvalues of g; these numbers lie on the unit circle $\mathbb{T}$ and only finitely many of them are distinct from 1. Any class function χ on $U(\infty)$ depends on the eigenvalues only, and we write it as $\chi(u_1, u_2, \dots)$.

Next, we need to introduce some notation. Let $\mathbb{R}_+ \subset \mathbb{R}$ denote the set of nonnegative real numbers, $\mathbb{R}_+^\infty$ denote the product of countably many copies of $\mathbb{R}_+$, and set

$$\mathbb{R}_+^{4\infty+2} = \mathbb{R}_+^\infty \times \mathbb{R}_+^\infty \times \mathbb{R}_+^\infty \times \mathbb{R}_+^\infty \times \mathbb{R}_+ \times \mathbb{R}_+.$$

Let $\Omega \subset \mathbb{R}_+^{4\infty+2}$ be the subset of sextuples

$$\omega = (\alpha^+, \beta^+; \alpha^-, \beta^-; \delta^+, \delta^-)$$

such that

$$\alpha^\pm = (\alpha_1^\pm \ge \alpha_2^\pm \ge \dots \ge 0) \in \mathbb{R}_+^\infty, \quad \beta^\pm = (\beta_1^\pm \ge \beta_2^\pm \ge \dots \ge 0) \in \mathbb{R}_+^\infty,$$
$$\sum_{i=1}^\infty (\alpha_i^\pm + \beta_i^\pm) \le \delta^\pm, \quad \beta_1^+ + \beta_1^- \le 1.$$

Equip $\mathbb{R}_+^{4\infty+2}$ with the product topology. An important fact is that, in the induced topology, Ω is a locally compact space.

Set

$$\gamma^{\pm} = \delta^{\pm} - \sum_{i=1}^{\infty} (\alpha_i^{\pm} + \beta_i^{\pm})$$

and note that γ^+, γ^- are nonnegative. For $u \in \mathbb{C}^*$ and $\omega \in \Omega$ set

$$\Phi(u;\omega) = e^{\gamma^+(u-1)+\gamma^-(u^{-1}-1)} \prod_{i=1}^{\infty} \frac{(1+\beta_i^+(u-1))(1+\beta_i^-(u^{-1}-1))}{(1-\alpha_i^+(u-1))(1-\alpha_i^-(u^{-1}-1))}. \tag{2.1}$$

For any fixed ω, this is a meromorphic function in variable $u \in \mathbb{C}^*$ with possible poles on $(0,1) \cup (1,+\infty)$. The poles do not accumulate to 1, so that the function is holomorphic in a neighborhood of $\mathbb{T}$.

Theorem 2.1. *The dual object $\widehat{U(\infty)}$ as defined above can be identified with the space Ω. More precisely, the extreme characters of the group $U(\infty)$ are the functions*

$$\chi_\omega(u_1, u_2, \dots) := \prod_{k=1}^{\infty} \Phi(u_k;\omega),$$

where ω ranges over Ω.

This is a deep result with a long history. See Voiculescu [54] and many references in Borodin-Olshanski [11, Section 1.1].

3. Relative dimension in the Gelfand-Tsetlin graph

The *Gelfand-Tsetlin graph* has the vertex set $\mathbb{S}_1 \sqcup \mathbb{S}_2 \sqcup \dots$ consisting of all signatures, and the edges formed by the couples (λ, ν) such that $\lambda \prec \nu$. This is a graded graph with the Nth level formed by $\mathbb{S}_N$.

By a *path* between two vertices $\varkappa \in \mathbb{S}_K$ and $\nu \in \mathbb{S}_N$, $K < N$, we mean a sequence

$$\varkappa = \lambda^{(K)} \prec \lambda^{(K+1)} \prec \dots \prec \lambda^{(N)} = \nu \in \mathbb{S}_N.$$

Such a path can be viewed as an array of numbers

$$\left\{\lambda_i^{(j)}\right\}, \quad K \le j \le N, \quad 1 \le i \le j,$$

satisfying the inequalities $\lambda_i^{(j+1)} \ge \lambda_i^{(j)} \ge \lambda_{i+1}^{(j+1)}$. It is called a *Gelfand–Tsetlin scheme*. If $K = 1$, the scheme has triangular form and if $K > 1$, it has trapezoidal form.

The triangular schemes with a fixed top level $\lambda^{(N)} = \nu$ parameterize the vectors of the *Gelfand-Tsetlin basis* in $\pi_\nu \in \widehat{U(N)}$, see Gelfand-Tsetlin [22], Zhelobenko [57]. The number of such schemes is equal to $\operatorname{Dim}_N \nu$.

The number of paths between $\varkappa$ and ν (or trapezoidal schemes with top ν and bottom $\varkappa$) will be denoted by $\operatorname{Dim}_{K,N}(\varkappa, \nu)$. It is equal to the quantity $\operatorname{Dim}(\pi_\varkappa, \pi_\nu)$ introduced in the preceding section.

Both $\operatorname{Dim}_N \nu$ and $\operatorname{Dim}_{K,N}(\varkappa, \nu)$ count the lattice points in some bounded convex polytopes.

Adding to the vertex set an additional 0th level formed by a singleton $\varnothing$, which is joined by edges with all vertices of level 1, one may write $\mathrm{Dim}_N\,\nu$ as $\mathrm{Dim}_{0,N}(\varnothing,\nu)$.

Note that the matrix Λ_K^N of format $\mathbb{S}_N \times \mathbb{S}_K$ that represents the link $\widehat{U(N)} \dashrightarrow \widehat{U(K)}$ coincides with the matrix product $\Lambda_{N-1}^N \dots \Lambda_K^{K+1}$, and its entries are given by

$$\Lambda_K^N(\nu,\varkappa) = \frac{\mathrm{Dim}_K\,\varkappa\,\mathrm{Dim}_{K,N}(\varkappa,\nu)}{\mathrm{Dim}_N\,\nu}.$$

A sequence of vertices $\{\lambda^{(N)} \in \mathbb{S}_N\}$ is said to be a *regular escape to infinity* if for every fixed vertex $\varkappa \in \mathbb{S}_K$ there exists a limit $\lim_{N\to\infty} \Lambda_K^N(\lambda^{(N)},\varkappa)$, and two regular escapes are called *equivalent* if the corresponding limits coincide for every $\varkappa$. The set of equivalence classes of regular escapes to infinity is called the *boundary* of the Gelfand-Tsetlin graph and denoted by $\partial\mathbb{GT}$. This is nothing else than one more, this time combinatorial, interpretation of the dual object $\widehat{U(\infty)}$.

Likewise, one can define the boundary $\partial\mathbb{Y}$ of the *Young graph*. That graph encodes the Young branching rule of the symmetric group characters, and $\partial\mathbb{Y}$ parameterizes the extreme characters of the infinite symmetric group.

In the symmetric group case, the stochastic links have the form

$$\Lambda_m^l(\lambda,\mu) = \begin{cases} \dfrac{\dim\mu\,\dim\lambda/\mu}{\dim\lambda}, & \text{if } \mu\subset\lambda,\\ 0, & \text{otherwise.}\end{cases}$$

where λ and μ are Young diagrams with l and m boxes, respectively ($l > m$), and $\dim(\,\cdot\,)$ denotes the number of standard Young tableaux of a given (possibly skew) shape.

As shown in Okounkov-Olshanski [39],

$$l(l-1)\dots(l-m+1)\frac{\dim\lambda/\mu}{\dim\lambda} = s_\mu^*(\lambda_1,\lambda_2,\dots), \tag{3.1}$$

where s_μ^* is the so-called *shifted Schur function*. Informally, the meaning of this result is that the quantity in the left-hand side behaves as a "regular" function in variable λ. Formula (3.1) is well suited for asymptotic analysis and makes it possible to quickly find the boundary $\partial\mathbb{Y}$, thus obtaining a proof of Thoma's theorem about the characters of the infinite symmetric group (Thoma [51]), see Kerov-Okounkov-Olshanski [26] and Borodin-Olshanski [13, Section 3.3].

By analogy, one can ask what can be said about the function

$$F_\varkappa(\nu) := \frac{\mathrm{Dim}_{K,N}(\varkappa,\nu)}{\mathrm{Dim}_N\,\nu}.$$

This problem was investigated in our recent paper Borodin-Olshanski [11]. To give a flavor of what we get, I will formulate one of the results in the simplest (but nontrivial!) case when $K = 1$.

Theorem 3.1. *Let $\varkappa = k$ range over $\mathbb{S}_1 = \mathbb{Z}$, ν range over $\mathbb{S}_N$, and write $F_k(\nu)$ instead of $F_\varkappa(\nu)$. Set*

$$H^*(t;\nu) = \prod_{j=1}^N \frac{t+j}{t+j-\nu_j},$$

where t is a formal variable.

Then the following identity holds

$$H^*(t;\nu)=\sum_{k\in\mathbb{Z}}F_k(\nu)\,\frac{(t+1)\dots(t+N)}{(t+1-k)\dots(t+N-k)}. \tag{3.2}$$

This is a kind of generating series for $F_k(\nu)$ from which one can extract a contour integral representation for $F_k(\nu)$.

Let $\varphi_n(\omega)$ denote the coefficients of the Laurent expansion of the function $u\mapsto\Phi(u;\omega)$ on $\mathbb{T}$:

$$\Phi(u;\omega)=\sum_{n\in\mathbb{Z}}\varphi_n(\omega)u^n. \tag{3.3}$$

The identity (3.2) mimics the Laurent expansion (3.3), and in a limit transition, (3.2) turns into (3.3).

I briefly list further results of [11].

There is an extension of (3.2) to arbitrary $K=1,2,\dots$ and $\varkappa\in\mathbb{S}_K$:

$$\prod_{i=1}^{K}H^*(t_i;\nu)=\sum_{\varkappa\in\mathbb{GT}_K}F_\varkappa(\nu)\,\mathfrak{S}_{\varkappa|N}(t_1,\dots,t_K), \tag{3.4}$$

where $\mathfrak{S}_{\varkappa|N}(t_1,\dots,t_K)$ is a certain "Schur-type" rational symmetric function in variables $t_1,\dots,t_K$:

$$\mathfrak{S}_{\varkappa|N}(t_1,\dots,t_K)=\operatorname{const}\,\frac{\det[G_{\varkappa_j+K-j|N}(t_i)]_{i,j=1}^{N}}{\prod_{1\le i<j\le N}(t_i-t_j)}$$

(here $G_{k|N}(t)$ are certain univariate rational functions).

We show that $F_\varkappa(\nu)$ also has a similarity with the Schur function. Namely, there is an analog of the Jacobi-Trudi formula:

$$F_\varkappa(\nu)=\det[F^{(j)}_{\varkappa_i-i+j}(\nu)]_{i,j=1}^{K},$$

where $F^{(j)}_k(\nu)$, $k\in\mathbb{Z}$, is a certain modification of $F_k(\nu)$. Note that a similar modified Jacobi-Trudi identity holds for the shifted Schur functions (Okounkov-Olshanski [39]) as well as for other analogs of Schur functions (Macdonald [35], Nakagawa-Noumi-Shirakawa-Yamada [36], Sergeev-Veselov [50]).

As both functions on the right-hand side of (3.4) are similar to the Schur functions, this relation may be viewed as a kind of the Cauchy identity.

From (3.2) one can derive a closed formula for $F_k(\nu)$ (in the form of a contour integral representation), and the same can be done for the modified functions $F^{(j)}_k(\nu)$. Like (3.1), the resulting formula is well adapted to asymptotic analysis, which enables us to re-derive Theorem 2.1 in a way very similar to that used in Kerov-Okounkov-Olshanski [26] for the infinite symmetric group $S(\infty)$.

Note that Petrov [49] found a different approach to the results of [11] together with a q-version of them.

Finally note that the results of [11] can also be extended to symplectic and orthogonal groups (work in progress).

4. The zw-measures

Let the symbol $\mathcal{P}(X)$ denote the set of probability measures on a space X. Given a measure $M \in \mathcal{P}(\mathbb{S}_N)$ with weights $M(\nu)$, its composition with the link Λ^N_{N-1} is a measure $M\Lambda^N_{N-1} \in \mathcal{P}(\mathbb{S}_{N-1})$ defined by

$$(M\Lambda^N_{N-1})(\lambda) = \sum_{\nu\in\mathbb{S}_N} M(\nu)\Lambda^N_{N-1}(\nu,\lambda), \qquad \lambda \in \mathbb{S}_{N-1}.$$

A family of measures $\{M_N \in \mathcal{P}(\mathbb{S}_N) : N = 1, 2, \dots\}$ is said to be *coherent* if the measures are consistent with the links in the sense that $M_N\Lambda^N_{N-1} = M_{N-1}$ for every $N \geq 2$.

For $\omega \in \Omega$ and $\nu \in \mathbb{S}_N$ we denote by $\Lambda^\infty_N(\omega,\nu)$ the coefficients in the expansion of the N-fold product $\Phi(u_1;\omega)\dots\Phi(u_N;\omega)$ on the normalized irreducible characters:

$$\begin{aligned}\Phi(u_1;\omega)\dots\Phi(u_N;\omega) &= \sum_{\nu\in\mathbb{S}_N} \Lambda^\infty_N(\omega,\nu)\widetilde{\chi}_\nu(u_1,\dots,u_N)\\ &= \sum_{\nu\in\mathbb{S}_N} \Lambda^\infty_N(\omega,\nu)\frac{\chi_\nu(u_1,\dots,u_N)}{\operatorname{Dim}_N\nu}.\end{aligned}$$

One readily shows that

$$\Lambda^\infty_N(\omega,\nu) = \operatorname{Dim}_N\nu\cdot\det[\varphi_{\nu_i-i+j}(\omega)]^N_{i,j=1}. \tag{4.1}$$

Note that Λ^∞_N is a Markov kernel meaning that for ω fixed, $\Lambda^\infty_N(\omega,\,\cdot\,)$ is a probability measure on $\mathbb{S}_N$. We regard Λ^∞_N as a "link" $\Omega \dashrightarrow \mathbb{S}_N$.

There is a natural one-to-one correspondence $\{M_N\} \leftrightarrow M_\infty$ between the coherent families $\{M_N\}$ and the measures $M_\infty \in \mathcal{P}(\Omega)$ given by

$$M_N = M_\infty\Lambda^\infty_N, \qquad N = 1, 2, 3, \dots.$$

Let us say that a quadruple z, z', w, w' of complex parameters is *admissible* if the following conditions hold: firstly, for every integer k, one has $(z+k)(z'+k) > 0$ and $(w+k)(w'+k) > 0$; secondly, $\Re(z+z'+w+w') > -1$. As readily seen, the first condition is equivalent to saying that each of pairs (z, z') and (w, w') belongs to the subset $\mathcal{Z} \subset \mathbb{C}^2$ defined by

$$\begin{aligned}\mathcal{Z} := \{(\zeta,\zeta') \in (\mathbb{C}\setminus\mathbb{Z})^2 \mid \zeta' = \bar{\zeta}\}\\ \cup\{(\zeta,\zeta') \in (\mathbb{R}\setminus\mathbb{Z})^2 \mid m < \zeta,\zeta' < m+1 \text{ for some } m \in \mathbb{Z}\}.\end{aligned} \tag{4.2}$$

For $N = 1, 2, \dots$ and $\nu \in \mathbb{S}_N$ set

$$\begin{aligned}M'_{z,z',w,w'|N}(\nu) = \prod_{i=1}^N \Bigg(&\frac{1}{\Gamma(z-\nu_i+i)\Gamma(z'-\nu_i+i)}\\ &\times\frac{1}{\Gamma(w+N+1+\nu_i-i)\Gamma(w'+N+1+\nu_i-i)}\Bigg)\cdot(\operatorname{Dim}_N\nu)^2.\end{aligned}$$

If (z, z', w, w') is admissible, then $M'_{z,z',w,w'|N}(\nu) > 0$, the series

$$\sum_{\nu\in\mathbb{S}_N} M'_{z,z',w,w'|N}(\nu)$$

is convergent, and its sum equals

$$C_{z,z',w,w'|N} := \prod_{i=1}^{N} \frac{\Gamma(z+z'+w+w'+i)}{\Gamma(z+w+i)\Gamma(z+w'+i)\Gamma(z'+w+i)\Gamma(z'+w'+i)\Gamma(i)}.$$

This is a multivariate version of Dougall's formula (1.3) we started with.

Consequently, the quantities

$$M_{z,z',w,w'|N}(\nu) := M'_{z,z',w,w'|N}(\nu)/C_{z,z',w,w'|N}, \qquad \nu \in \mathbb{S}_N,$$

determine a probability measure on $\mathbb{S}_N$. For $N = 1$ this measure coincides with the measure $M_{z,z',w,w'}$ on $\mathbb{Z}$ introduced in the very beginning.

The measures $M_{z,z',w,w'|N}$ are a special case of the *orthogonal polynomial ensembles* (about this notion see König [30]).

Namely, let us associate with $\nu \in \mathbb{S}_N$ a collection $(n_1, \dots, n_N)$ of pairwise distinct integers by setting

$$n_i := \nu_i + N - i, \qquad i = 1, \dots, N. \tag{4.3}$$

Under the correspondence $\nu \mapsto (n_1, \dots, n_N)$, the measure $M_{z,z',w,w'|N}$ determines an ensemble of random N-point configurations on $\mathbb{Z}$, and it is the orthogonal polynomial ensemble related to the family of polynomials orthogonal with respect to weight $M_{z+N-1,z'+N-1,w,w'}$. These curious orthogonal polynomials were discovered by Askey [1] and later investigated by Lesky [31, 32]. They are relatives of the classical Hahn polynomials. For more detail, see Borodin-Olshanski [12].

Theorem 4.1. *Fix a quadruple (z, z', w, w') of admissible parameters and let N range over $\{1, 2, \dots\}$. The family $\{M_{z,z',w,w'|N}(\nu)\}$ just defined is a coherent family.*

Different proofs are given in Olshanski [42, 43]. The latter paper actually contains a more general result (the links and the measures depend on an additional parameter — the "Jack parameter"). In Olshanski-Osinenko [48], the theorem is extended to other branching graphs including those related to the orthogonal and symplectic groups.

Corollary 4.2. *For every admissible quadruple (z, z', w, w') there exists a probability measure $M_{z,z',w,w'|\infty}$ on the space Ω, uniquely determined by the property that*

$$M_{z,z',w,w'|\infty}\Lambda_N^\infty = M_{z,z',w,w'|N}, \qquad N = 1, 2, \dots .$$

Both $M_{z,z',w,w'|N}$ and $M_{z,z',w,w'|\infty}$ are called the *zw-measures*. They are analogs of the *z-measures* which arise in the context of the symmetric groups (see Borodin-Olshanski [8], the recent survey Olshanski [44], and also Section 7 below). A common feature of all these measures is that they serve as the laws of *determinantal point processes* (about those, see Borodin [4] and references therein).

It is worth noting that the zw-measures on $\mathbb{S}_N$ are introduced by an explicit formula while our definition of the zw-measures on Ω is indirect: Corollary 4.2 only provides the explicit values for the expectation of certain functionals.

Our interest in the zw-measures on Ω is motivated by the fact that they arise in the problem of harmonic analysis on the infinite-dimensional unitary group (Olshanski [42], Borodin-Olshanski [9, 10]).

5. Markov processes

We need a few basic notions from the theory of Markov processes (see Liggett [33], Ethier-Kurtz [21]).

Let X be a locally compact space and $C_0(X)$ denote the space of real-valued continuous functions on X vanishing at infinity; $C_0(X)$ is a Banach space with respect to the supremum norm. A *Feller semigroup* on X is a strongly continuous semigroup $P(t)$, $t \geq 0$, of operators on $C_0(X)$ which are given by Markov kernels. This means that

$$(P(t)f) = \int_X P(t; x, dy) f(y), \qquad x \in X, \quad f \in C_0(X),$$

where $P(t; x, \cdot) \in \mathcal{P}(X)$ for every $t \geq 0$ and $x \in X$. A well-known abstract theorem says that a Feller semigroup gives rise to a Markov process on X with transition function $P(t; x, dy)$. The processes derived from Feller semigroups are called *Feller processes*; they form a particularly nice subclass of general Markov processes.

A Feller semigroup $P(t)$ is uniquely determined by its *infinitesimal generator*. This is a (typically, unbounded) closed operator A on $C_0(X)$ given by

$$Af = \lim_{t \to +0} \frac{P(t)f - f}{t}.$$

The *domain* of A, denoted by $\operatorname{Dom} A$, is the (algebraic) subspace formed by those functions $f \in C_0(X)$ for which the above limit exists; $\operatorname{Dom} A$ is always a dense subspace. A *core* of A is a subspace $\mathcal{F} \subset \operatorname{Dom} A$ such that A is the closure of $A\big|_{\mathcal{F}}$. One can say that a core is an "essential domain" for A. The full domain is often difficult to describe explicitly, and then one is satisfied by exhibiting a core with the action of A on it.

The Markov chain on $X = \mathbb{Z}$ mentioned in Section 1 is an example of a Feller process. Now we are going to define its multidimensional analog with $X = \mathbb{S}_N$.

First we need to introduce some notation. It is convenient to use the correspondence (4.3) to pass from $\mathbb{S}_N$ to the subset $\Omega_N \subset \mathbb{Z}^N$ formed by the N-tuples $\widetilde{n} = (n_1 > \cdots > n_N)$. Let

$$V(\widetilde{n}) := \prod_{1 \leq i < j \leq N} (n_i - n_j)$$

and ε_k denote the kth basis vector in $\mathbb{Z}^N \subset \mathbb{R}^N$.

We introduce a partial difference operator $D_{z,z',w,w'|N}$ on Ω_N depending on an admissible quadruple (z, z', w, w'), as follows

$$\begin{aligned}
&(D_{z,z',w,w'|N} f)(\widetilde{n}) \\
&\quad = \sum_{k=1}^{N} \Bigg(\frac{V(\widetilde{n} + \varepsilon_k)}{V(\widetilde{n})} (z + N - 1 - n_k)(z' + N - 1 - n_k)(f(\widetilde{n} + \varepsilon_k) - f(\widetilde{n})) \\
&\qquad\qquad + \frac{V(\widetilde{n} - \varepsilon_k)}{V(\widetilde{n})} (w + n_k)(w' + n_k)(f(\widetilde{n} - \varepsilon_k) - f(\widetilde{n})) \Bigg) + \text{const}, \quad (5.1)
\end{aligned}$$

where the constant term is chosen so that the operator annihilates the constant functions.

This difference operator is well defined on Ω_N, because if $\widetilde{n} + \varepsilon_k \notin \Omega_N$, or $\widetilde{n} - \varepsilon_k \notin \Omega_N$, then $V(\widetilde{n} + \varepsilon_k)$ or, respectively, $V(\widetilde{n} - \varepsilon_k)$ vanishes.

In the case $N = 1$ the operator reduces to the ordinary difference operator $D_{z,z',w,w'}$ defined in (1.4).

Theorem 5.1. *Let (z, z', w, w') be an admissible quadruple of parameters. For every $N = 1, 2, \dots$ there exists a Feller semigroup on $\Omega_N \subset \mathbb{Z}^N$ whose generator is given by the partial difference operator* (5.1) *with domain*

$$\{f \in C_0(\Omega_N) : D_{z,z',w,w'|N} f \in C_0(\Omega_N)\}.$$

See Borodin-Olshanski [12, Section 5].

As pointed out in [12, Subsection 1.3], the Feller process provided by Theorem 5.1 can be viewed as the *Doob h-transform* of a collection of N independent Markov chains on $\mathbb{Z}$, with h equal to the Vandermonde $V(\widetilde{n})$.

Using the bijection (4.3) between $\mathbb{S}_N$ and Ω_N we may interpret the semigroup of Theorem 5.1 as a Feller semigroup on $C_0(\mathbb{S}_N)$. Let us denote it by $P_{z,z',w,w'|N}(t)$.

Theorem 5.2. *The measure $M_{z,z',w,w'|N}$ on $\mathbb{S}_N$ serves as the stationary distribution for the Feller process defined by the semigroup $P_{z,z',w,w'|N}(t)$.*

See Borodin-Olshanski [12, Section 7].

Theorem 5.3. *Let (z, z', w, w') be a fixed admissible quadruple and N range over $\{1, 2, \dots\}$. The links Λ^N_{N-1} intertwine the semigroups $P_{z,z',w,w'|N}(t)$.*

See Borodin-Olshanski [12, Section 6]. Let us comment on this result. The link Λ^N_{N-1} determines an operator $f \mapsto \Lambda^N_{N-1} f$ transforming bounded functions on $\mathbb{S}_{N-1}$ into bounded functions on $\mathbb{S}_N$ by

$$(\Lambda^N_{N-1} f)(\nu) = \sum_{\lambda \in \mathbb{S}_{N-1}} \Lambda^N_{N-1}(\nu, \lambda) f(\lambda).$$

One proves that Λ^N_{N-1} is "Feller" in the sense that it maps $C_0(\mathbb{S}_{N-1})$ into $C_0(\mathbb{S}_N)$, and the claim of the theorem means that the following commutativity relations hold

$$P_{z,z',w,w'|N}(t)\Lambda^N_{N-1} = \Lambda^N_{N-1} P_{z,z',w,w'|N-1}(t), \qquad N = 2, 3, \dots, \quad t \geq 0.$$

One also proves the Feller property for the link Λ^∞_N meaning that Λ^∞_N maps $C_0(\mathbb{S}_N)$ into $C_0(\Omega)$. (Because of (4.1), this amounts to the fact that the functions $\varphi_n(\omega)$ lie in $C_0(\Omega)$.) Then, using a very simple argument, one derives from the above theorems the following result:

Theorem 5.4.

(i) *For every admissible quadruple (z, z', w, w'), there exists a unique Feller semigroup $P_{z,z',w,w'|\infty}(t)$ on $C_0(\Omega)$ such that*

$$P_{z,z',w,w'|\infty}(t)\Lambda^\infty_N = \Lambda^\infty_N P_{z,z',w,w'|N}(t), \qquad N = 1, 2, 3, \dots, \quad t \geq 0.$$

(ii) *The measure $M_{z,z',w,w'|\infty}$ on Ω serves as the stationary distribution for the corresponding Feller process.*

This is one of the main results of Borodin-Olshanski [12] (see also the expository paper Olshanski [46]).

6. The representation ring and the generator

In this section I briefly review the recent results from my paper [47].

Let R^S denote the graded representation ring of all symmetric groups $S(n)$ collected together, with the multiplication determined by the operation of induction $\mathrm{Ind}_{S(m)\times S(n)}^{S(m+n)}$ from Young subgroups: see Macdonald [34, Chapter I, Section 7]. As is clearly shown there, the original Frobenius' approach to the classification of the symmetric group characters essentially relies on the canonical isomorphism between R^S and the ring of symmetric functions (see also Zelevinsky [56]).

One can ask if there is a reasonable analog of the ring R^S for the unitary groups (as well as for other families of classical compact groups). The answer is yes, but it is necessary to take into account the fact that the operation of induction $\mathrm{Ind}_{U(M)\times U(N)}^{U(M+N)}$ leads to *infinite* sums of irreducible representations.

Let $\mathbb{C}[[\dots,\varphi_{-1},\varphi_0,\varphi_1,\dots]]$ be the $\mathbb{C}$-algebra of formal power series in countably many indeterminates φ_n, $n\in\mathbb{Z}$, and let

$$\mathbb{C}[[\dots,\varphi_{-1},\varphi_0,\varphi_1,\dots]]_{\mathrm{bounded}}\subset\mathbb{C}[[\dots,\varphi_{-1},\varphi_0,\varphi_1,\dots]]$$

be the subalgebra of series with bounded degree. This subalgebra is a graded algebra: its Nth homogeneous component is formed by the homogeneous series of degree N.

According to our definition, the representation ring for the family of unitary groups, denoted by R^U, can be identified with $\mathbb{C}[[\dots,\varphi_{-1},\varphi_0,\varphi_1,\dots]]_{\mathrm{bounded}}$.

The algebra R^U contains all the irreducible characters χ_ν (where $\nu\in\mathbb{S}_N$, $N=1,2,\dots$): namely we identify

$$\chi_\nu=\det[\varphi_{\nu_i-i+j}]_{i,j=1}^N. \tag{6.1}$$

We introduce an "adic" topology in R^U. With respect to it, both the monomials in the indeterminates φ_n and the characters χ_ν (together with the unity element 1) are "topological bases".

Now let us fix an arbitrary quadruple (z,z',w,w') of complex parameters and introduce the following formal differential operator in countably many indeterminates $\{\varphi_n : n\in\mathbb{Z}\}$

$$\mathbb{D}_{z,z',w,w'}=\sum_{n\in\mathbb{Z}}A_{nn}\frac{\partial^2}{\partial\varphi_n^2}+2\sum_{\substack{n_1,n_2\in\mathbb{Z}\\ n_1>n_2}}A_{n_1n_2}\frac{\partial^2}{\partial\varphi_{n_1}\partial\varphi_{n_2}}+\sum_{n\in\mathbb{Z}}B_n\frac{\partial}{\partial\varphi_n},$$

where, for any indices $n_1\ge n_2$,

$$A_{n_1n_2}=\sum_{p=0}^{\infty}(n_1-n_2+2p+1)(\varphi_{n_1+p+1}\varphi_{n_2-p}+\varphi_{n_1+p}\varphi_{n_2-p-1})$$
$$-(n_1-n_2)\varphi_{n_1}\varphi_{n_2}-2\sum_{p=1}^{\infty}(n_1-n_2+2p)\varphi_{n_1+p}\varphi_{n_2-p}$$

and, for any $n\in\mathbb{Z}$,

$$B_n=(n+w+1)(n+w'+1)\varphi_{n+1}+(n-z-1)(n-z'-1)\varphi_{n-1}$$
$$-\big((n-z)(n-z')+(n+w)(n+w')\big)\varphi_n.$$

The operator $\mathbb{D}_{z,z',w,w'}$ correctly defines a linear map $R^U \to R^U$. Notice that only the coefficients B_n depend on the parameters (z, z', w, w').

Our aim is to interpret $\mathbb{D}_{z,z',w,w'}$ as an operator acting on a certain linear subspace $\mathcal{F} \subset C_0(\Omega)$.

First, we define $\mathcal{F}$ as the algebraic linear span of all the elements $\chi_\nu \in R^U$ (where $\nu \in \mathbb{S}_N$, $N = 1, 2, \dots$).

Proposition 6.1. *The subspace $\mathcal{F}$ is invariant under the action of operator $\mathbb{D}_{z,z',w,w'}$.*

Next, we embed $\mathcal{F}$ into $C_0(\Omega)$. To this end we identify every formal indeterminate φ_n with the function $\varphi_n(\omega)$ on Ω introduced in (3.3). (We have already mentioned that these functions lie in $C_0(\Omega)$.) Then, by virtue of (6.1), all elements $\chi_\nu \in R^R$ are turned into functions $\chi_\nu(\omega)$ belonging to $C_0(\Omega)$. In this way $\mathcal{F}$ becomes a subspace of $C_0(\Omega)$.

In the next theorem we use the notion of a core defined in the beginning of Section 5.

Theorem 6.2. *Assume (z, z', w, w') is admissible and let $A_{z,z',w,w'}$ denote the generator of the Feller semigroup $P_{z,z',w,w'|\infty}(t)$ from Theorem 5.4.*

The subspace $\mathcal{F} \subset C_0(\Omega)$ is an invariant core for the generator $A_{z,z',w,w'}$, and its action on $\mathcal{F}$ is implemented by the operator $\mathbb{D}_{z,z',w,w'}$.

Our construction of the Feller processes on Ω is rather abstract and indirect, but Theorem 6.2 provides a piece of concrete information about them.

7. The Young bouquet

Here I review the results of Borodin-Olshanski [13].

Consider the infinite chain of finite symmetric groups with natural embeddings

$$S(1) \subset S(2) \subset S(3) \subset \dots \tag{7.1}$$

and let $S(\infty)$ denote the union of all these groups. In other words, $S(\infty)$ is the group of finitary permutations of the set $\{1, 2, 3, \dots\}$. The characters of both the symmetric and unitary groups are related to the Schur symmetric functions. The similarity between the characters of the inductive limit groups $S(\infty)$ and $U(\infty)$ is even more apparent. On the other hand, the groups themselves are structurally very different. We suggest an explanation of this phenomenon.

As we tried to demonstrate in Section 3, the combinatorial base of the character theory of $U(\infty)$ is the Gelfand-Tsetlin graph. Its counterpart in the symmetric group case is the *Young graph*, also called the *Young lattice*. The vertex set of the Young graph is the set of all Young diagrams, and two diagrams are joined by an edge if they differ by a single box. The graph is graded: its nth level ($n = 0, 1, 2, \dots$) consists of the diagrams with n boxes. The Young graph encodes the branching of the irreducible characters of the chain (7.1), just as the Gelfand-Tsetlin graph does for the chain (1.1) (Vershik-Kerov [52]).

The description of the extreme characters of $S(\infty)$ was given by Thoma [51]. It can be reformulated as the description of the boundary of the Young graph. For various proofs of the fundamental Thoma's theorem, see Vershik-Kerov [52], Okounkov [37], Kerov-Okounkov-Olshanski [26].

In [13] we introduce and study a new object which occupies an intermediate position between the Gelfand-Tsetlin graph and the Young graph, and makes it possible to see a clear

connection between them. This new object is called the *Young bouquet* and denoted as $\mathbb{YB}$. It is not an ordinary graph. However, $\mathbb{YB}$ is a graded poset, similarly to the Gelfand-Tsetlin and Young graphs.

One new feature is that the grading in $\mathbb{YB}$ is not discrete but continuous: the grading level is marked by a positive real number. By definition, the elements of $\mathbb{YB}$ of a given level $r > 0$ are pairs (λ, r), where λ is an arbitrary Young diagram. The partial order in $\mathbb{YB}$ is defined as follows: $(\mu, r) < (\lambda, r')$ if $r < r'$ and diagram μ is contained in diagram λ (or coincides with it).

From the definition of the Young bouquet one sees that it is closely related to the Young graph. We explain in [13] how various notions related to the Young graph are modified in the context of the Young bouquet. In particular, we are led to consider *Young tableaux with continuous entries* as a counterpart of the conventional tableaux.

Let $\mathbb{YB}_r$ stand for the rth level of the poset $\mathbb{YB}$; this is simply a copy of the set $\mathbb{Y}$ of all Young diagrams. For every couple of positive reals $r' > r$ we define a link ${}^{\mathbb{YB}}\Lambda_r^{r'} : \mathbb{YB}_{r'} \dashrightarrow \mathbb{YB}_r$, which is a stochastic matrix of format $\mathbb{Y} \times \mathbb{Y}$ that depends only of the ratio $r' : r$. The links satisfy the compatibility relation

$$ {}^{\mathbb{YB}}\Lambda_{r'}^{r''} \, {}^{\mathbb{YB}}\Lambda_r^{r'} = {}^{\mathbb{YB}}\Lambda_r^{r''}, \qquad r'' > r' > r > 0, $$

which enables us to define the *boundary of the Young bouquet* in the spirit of the fourth version of the definition given in Section 2.

The boundary of $\mathbb{YB}$ and the boundary of the Young graph are directly connected: the former is the cone whose base is the latter. Namely, the boundary of $\mathbb{YB}$, called the *Thoma cone*, can be identified with the subset in $\mathbb{R}_+^\infty \times \mathbb{R}_+^\infty \times \mathbb{R}_+$ formed by the triples (α, β, δ) such that

$$ \alpha = (\alpha_1 \geq \alpha_2 \geq \cdots \geq 0), \quad \beta = (\beta_1 \geq \beta_2 \geq \cdots \geq 0), \quad \delta \geq 0 $$

and

$$ \sum_{i=1}^{\infty} (\alpha_i + \beta_i) \leq \delta, $$

while the boundary of the Young graph, called the *Thoma simplex*, can be identified with the section $\delta = 1$ of the Thoma cone.

On the other hand, we explain how the Young bouquet $\mathbb{YB}$ is connected with the Gelfand-Tsetlin graph. We consider the subgraph $\mathbb{GT}^+$ of the Gelfand-Tsetlin graph spanned by the signatures with nonnegative coordinates. The Nth level vertices of $\mathbb{GT}^+$ can be viewed as pairs (λ, N), where λ is a Young diagram with at most N nonzero rows.

We show (Theorem 4.4.1 in [13]) that $\mathbb{YB}$ is a *degeneration* of $\mathbb{GT}^+$ in the following sense.

Theorem 7.1. *Fix arbitrary positive numbers $r' > r > 0$ and arbitrary two Young diagrams λ and μ such that $\mu \subseteq \lambda$. Let two positive integers $N' > N$ go to infinity in such a way that $N'/N \to r'/r$. Then*

$$ \lim \Lambda_N^{N'}((\lambda, N'), (\mu, N)) = {}^{\mathbb{YB}}\Lambda_r^{r'}(\lambda, \mu). \tag{7.2} $$

We also exhibit a limit procedure turning the boundary of $\mathbb{GT}^+$ (which is a subset of Ω) into the boundary of $\mathbb{YB}$, the Thoma cone (Theorem 4.5.1 in [13]).

Next, we show that along the degeneration $\mathbb{GT}^+ \to \mathbb{YB}$, some degenerate versions of zw-measures on the levels of the Gelfand-Tsetlin graph turn into the *z-measures* on the set $\mathbb{Y}$.

The z-measures on $\mathbb{Y}$ are a distinguished particular case of Okounkov's *Schur measures* (Okounkov [38], Borodin-Okounkov [7]). For the first time, the z-measures appeared in Borodin-Olshanski [8] in connection with the problem of harmonic analysis on the infinite symmetric group stated in Kerov-Olshanski-Vershik[27] (see also the detailed paper [28]).

The z-measures depend on a pair $(z, z') \in \mathcal{Z}$ of parameters (see (4.2)) and the additional parameter $r > 0$ indexing the level of $\mathbb{YB}$. The measures are consistent with the links ${}^{\mathbb{YB}}\Lambda_r^{r'}$ and give rise to certain probability measures $M_{z,z'|\infty}$ on the Thoma cone, in complete similarity with the case of Gelfand-Tsetlin graph (see Corollary 4.2 above).

The parallelism between the Young bouquet and the Gelfand-Tsetlin graph also extends to the theory outlined in Section 5. In Borodin-Olshanski [14] we show that using the same approach, one can construct a family of Feller Markov processes on the Thoma cone.

8. Notes and complements

8.1. In connection with the material of Section 3 see also Olshanski [45].

8.2. There exist other values of parameters (z, z', w, w') for which coherent families $\{M_{z,z',w,w'|N}; N = 1, 2, \dots\}$ are still well defined and give rise to certain probability measures $M_{z,z',w,w'|\infty}$ on the boundary Ω. Only these measures are *degenerate* meaning that the support of $M_{z,z',w,w'|N}$ is a proper subset of $\mathbb{S}_N$.

For instance, one can take

$$z = m, \quad z' = m + a, \quad w = 0, \quad w' = b,$$

where m is a positive integer, $a > -1$, and $b > -1$. Then the corresponding measure on Ω is concentrated on the subset

$$\{\omega : \ \delta^+ = \beta_1^+ + \cdots + \beta_m^+, \quad 1 \geq \beta_1^+ \geq \cdots \geq \beta_m^+ \geq 0, \\ \text{all other coordinates equal } 0\} \subset \Omega$$

and takes the form

$$(\text{normalizing constant}) \cdot \prod_{i=1}^{m} t_i^b (1 - t_i)^a \cdot \prod_{1 \leq i < j \leq m} (t_i - t_j)^2 \cdot \prod_{i=1}^{m} dt_i, \tag{8.1}$$

where we use the notation

$$(t_1, \dots, t_m) := (1 - \beta_m^+, \dots, 1 - \beta_1^+).$$

The measure (8.1) is a multidimensional version of the Euler Beta distribution of the type appearing in Selberg's integral, and the coherency property of the degenerate zw-measures is related to a generalized Selberg integral (Olshanski [43, Section 5]).

8.3. In the case of degenerate measures, the construction of Section 6 produces a diffusion Markov process on the m-dimensional simplex

$$\{(t_1,\dots,t_m): 1\ge t_1\ge\dots\ge t_m\ge 0\}$$

whose generator is the *m-variate Jacobi differential operator*

$$D_m^{(a,b)} := \frac{1}{V_m}\circ\left(\sum_{i=1}^m\left(t_i(1-t_i)\frac{\partial^2}{\partial t_i^2}+[b+1-(a+b+2)t_i]\frac{\partial}{\partial t_i}\right)\right)\circ V_m+\text{const}$$
$$=\sum_{i=1}^m\left(t_i(1-t_i)\frac{\partial^2}{\partial t_i^2}+\left[b+1-(a+b+2)t_i+\sum_{j:\,j\ne i}\frac{2t_i(1-t_i)}{t_i-t_j}\right]\frac{\partial}{\partial t_i}\right),$$

where V_m denotes the Vandermonde,

$$V_m := \prod_{1\le i<j\le m}(t_i-t_j),$$

and

$$\text{const}=\sum_{k=0}^{m-1}k(k+a+b+1).$$

This fact is substantially exploited in the proof of Theorem 6.2.

The same diffusion process also arises in a different context, see Gorin [23].

8.4. Gorin [25] considered the "q-Gelfand-Tsetlin graph" (which amounts to introducing a q-deformation of the links $\Lambda_{N-1}^N : \mathbb{S}_N \dashrightarrow \mathbb{S}_{N-1}$) and found the corresponding boundary. Under this deformation, the α-parameters disappear while the β-parameters survive but become discrete.

For the "q-Gelfand-Tsetlin graph", analogs of zw-measures and related processes are unknown. However, Borodin and Gorin [6] applied the approach outlined in Section 6 above to constructing Feller processes of a different kind.

8.5. Let $\mathcal{T}$ stand for the space of infinite monotone paths in the Gelfand-Tsetlin graph, which are the same as *infinite Gelfand-Tsetlin schemes*. The path space $\mathcal{T}$ plays an important role in our theory.

There exists a one-to-one correspondence $\mathcal{P}(\Omega)\leftrightarrow\mathcal{P}_{\text{central}}(\mathcal{T})$ between probability measures on Ω and some kind of *Gibbs measures* (or *central measures*, in Vershik-Kerov's terminology) on $\mathcal{T}$ (Olshanski [42, Proposition 10.3]).

Using this correspondence, Gorin [24] proved that the zw-measures on Ω are pairwise mutually singular.

Via this correspondence, the semigroup $P_{z,z',w,w'|\infty}(t)$ defines an evolution of central measures. It is natural to ask if there exists a Markov process on $\mathcal{T}$ that agrees with that evolution when restricted to the central measures. In Borodin-Olshanski [12] we construct one such process (for every admissible (z,z',w,w')).

In Borodin-Olshanski [15] we present arguments in favor of the conjecture that a similar construction can be carried out for the Young bouquet. (Then the path space consists of infinite Young tableaux with continuous entries.) The conjectural process should be *piecewise deterministic* meaning that it is a combination of a dynamical system with a jump Markov process.

8.6. Note that in the literature one can find a number of examples of "Markov intertwiners" (that is, Markov kernels intertwining two Markov processes); see, e.g. Biane [2], [3], Dubédat [18], Carmona-Petit-Yor[16]. However, the use of Markov intertwiners for constructing infinite-dimensional Markov processes seems to be new.

8.7. In the ICM lecture [5], Borodin demonstrates how intertwined Markov processes of the type considered above are applied to analyzing the large time behavior of certain interacting particle systems and random growth models.

Acknowledgements. The author was partially supported by the Simons Foundation and the RFBR grant 13-01-12449.

References

[1] Askey, R., *An integral of Ramanujan and orthogonal polynomials*, J. Indian Math. Soc **51** (1987), 27–36.

[2] Biane, Ph., *Intertwining of Markov semigroups, some examples*, Séminaire de Probabilités XXIX, 30–36, Lecture Notes in Math., **1613**, Springer, Berlin, 1995.

[3] ———, *Entrelacements de semi-groupes provenant de paires de Gelfand*, ESAIM Probab. Stat. **15** (2011), In honor of Marc Yor, suppl., S2–S10.

[4] Borodin, A., *Determinantal point processes*, In: The Oxford Handbook on Random Matrix Theory, Gernot Akemann, Jinho Baik, and Philippe Di Francesco, eds. Oxford University Press, 2011, Chapter 11, 231-249; arXiv:0911.1153.

[5] ———, *Integrable probability*, Proceedings of ICM-2014, Seoul.

[6] Borodin, A. and Gorin V., *Markov processes of infinitely many nonintersecting random walks*, Probab. Theory Related Fields **155** (2013), no. 3-4, 935–997; arXiv:1106.1299.

[7] Borodin, A. and Okounkov, A., *A Fredholm determinant formula for Toeplitz determinants*, Integral Equations Operator Theory **37** (2000), no. 4, 386–396; arXiv:math/9907165.

[8] Borodin, A. and Olshanski, G., *Distributions on partitions, point processes, and the hypergeometric kernel*, Commun. Math. Phys. **211** (2000), no. 2, 335–358; arXiv:math/9904010.

[9] ———, *Harmonic analysis on the infinite–dimensional unitary group and determinantal point processes*, Ann. Math. **161** (2005), no.3, 1319–1422; arXiv:math/0109194.

[10] ———, *Representation theory and random point processes*, In: A. Laptev (ed.), European congress of mathematics, Stockholm, Sweden, June 27–July 2, 2004. Zürich: European Mathematical Society, 2005, pp. 73–94; arXiv:math/0409333.

[11] ———, *The boundary of the Gelfand-Tsetlin graph: A new approach*, Advances in Math. **230** (2012), 1738–1779; arXiv:1109.1412.

[12] ———, *Markov processes on the path space of the Gelfand-Tsetlin graph and on its boundary*, Journal of Functional Analysis **263** (2012), 248–303; arXiv:1009.2029.

[13] ———, *The Young bouquet and its boundary*, Moscow Mathematical Journal **13** (2013), no. 2, 191–230; arXiv:1110.4458.

[14] ———, *Markov dynamics on the Thoma cone: a model of time-dependent determinantal processes with infinitely many particles*, Electronic Journal of Probability **18** (2013), no. 75, 1–43; arXiv:1303.2794.

[15] ———, *An interacting particle process related to Young tableaux*, Zapiski Nauchnyh Seminarov POMI [Proceedings of the Scientific Seminars of the Steklov Mathematical Institute, St.-Petersburg Branch], vol. 421 (2014), 47–57 [to be reproduced in J. Math. Sci. (New York)]; arXiv:1303.2795.

[16] Carmona, P., Petit, F., and Yor, M., *Beta-gamma random variables and intertwining relations between certain Markov processes*, Revista Matemática Iberoamericana **14** (1998), No 2, 311–367.

[17] Dougall, J., *On Vandermonde's theorem and some more general expansions*, Proc. Edinburgh Math. Soc. **25** (1907), 114–132.

[18] Dubédat, J., *Reflected planar Brownian motions, intertwining relations and crossing probabilities*, Annales Institut Henri Poincaré, Sér. Prob. Stat. **40** (2004), 539–552; arXiv:math/0302250.

[19] Edrei, A., *On the generating function of a doubly infinite, totally positive sequence*, Trans. Amer. Math. Soc. **74** (1953), 367–383.

[20] Erdelyi, A. (editor), *Higher transcendental functions*, vol. I. McGraw–Hill, New York, 1953.

[21] Ethier, S. N. and Kurtz, T. G., *Markov processes — Characterization and convergence*, Wiley–Interscience, New York 1986.

[22] Gelfand, I. M. and Tsetlin, M. L., *Finite-dimensional representations of the group of unimodular matrices*, Doklady Akad. Nauk SSSR **71** (1950), 825-828 (Russian); English translation: I. M. Gelfand, Collected papers, vol. 2, Springer, 1988.

[23] Gorin, V. E., *Non-colliding Jacobi diffusions as the limit of Markov chains on the Gelfand-Tsetlin graph*, Zapiski Nauchnyh Seminarov POMI [Proceedings of the Scientific Seminars of the Steklov Mathematical Institute, St.-Petersburg Branch] **360** (2008), 91–123; translation in J. Math. Sci. (N. Y.) **158** (2009), no. 6, 819–837; arXiv:0812.3146.

[24] ———, *Disjointness of representations arising in the problem of harmonic analysis on an infinite-dimensional unitary group*, Funktsional. Anal. i Prilozhen. **44** (2010), no. 2, 14–32 (Russian); translation in Funct. Anal. Appl. **44** (2010), no. 2, 92–105; arXiv:0805.2660.

[25] Gorin, V., *The q-Gelfand-Tsetlin graph, Gibbs measures and q-Toeplitz matrices*, Advances in Math. **229** (2012), no. 1, 201–266; arXiv:1011.1769.

[26] Kerov, S., Okounkov, A., and Olshanski, G., *The boundary of Young graph with Jack edge multiplicities*, Intern. Mathematics Research Notices, 1998, no. 4, 173–199; arXiv:q-alg/9703037.

[27] Kerov, S., Olshanski, G., and Vershik, A., *Harmonic analysis on the infinite symmetric group. A deformation of the regular representation*, Comptes Rendus Acad. Sci. Paris. Sér. 1, 316 (1993), 773–778.

[28] ———, *Harmonic analysis on the infinite symmetric group*, Inventiones Mathematicae **158** (2004), no. 3, 551–642; arXiv:math/0312270.

[29] Kirillov, A. A., *Elements of the theory of representations*, Grundlehren der mathematischen Wissenschaften **220**, Springer, 1976.

[30] König, W., *Orthogonal polynomial ensembles in probability theory*, Probab. Surveys **2** (2005), 385–447; arXiv:math/0403090.

[31] Lesky, P. A., *Unendliche und endliche Orthogonalsysteme von continuous Hahnpolynomen*, Results Math. **31** (1997), 127–135.

[32] ———, *Eine Charakterisierung der kontinuierlichen und diskreten klassischen Orthogonalpolynome*, Preprint 9812, Mathematisches Institut A, Universität Stuttgart.

[33] Liggett, T. M., *Continuous time Markov processes*, Graduate Texts in Math. 113. Amer. Math. Soc., 2010.

[34] Macdonald, I. G., *Symmetric functions and Hall polynomials*, 2nd edition, Oxford University Press, 1995.

[35] ———, *Schur functions: theme and variations*, In: Séminaire Lotharingien de Combinatoire **28** (1992), 35 pp.

[36] Nakagawa, J., Noumi, M., Shirakawa, M., and Yamada, Y., *Tableau representation for Macdonald's ninth variation of Schur functions*, In: Physics and combinatorics, 2000 (Nagoya), 180–195, World Sci. Publ., River Edge, NJ, 2001.

[37] Okounkov, A., *On representations of the infinite symmetric group*, Zapiski Nauchnyh Seminarov POMI [Proceedings of the Scientific Seminars of the Steklov Mathematical Institute, St.-Petersburg Branch] **240** (1997), 166–228, (Russian); translation in J. Math. Sci. (New York) **96** (1999), no. 5, 3550–3589; arXiv:math/9803037.

[38] ———, *Infinite wedge and random partitions*, Selecta Math. (N.S.) **7** (2001), no. 1, 57–81; arXiv:math/9907127.

[39] Okounkov, A. and Olshanski, G., *Shifted Schur functions*, Algebra i Analiz **9** (1997), no. 2, 73–146 (Russian); English version: St. Petersburg Mathematical J. **9** (1998), 239–300; arXiv:q-alg/9605042.

[40] ———, *Asymptotics of Jack polynomials as the number of variables goes to infinity*, Intern. Math. Research Notices **1998** (1998), no. 13, 641–682; arXiv:q-alg/9709011.

[41] Olshanski, G., *Unitary representations of infinite-dimensional pairs* (G, K) *and the formalism of R. Howe*, In: Representations of Lie groups and related topics. Advances in Contemp. Math., vol. 7 (A. M. Vershik and D. P. Zhelobenko, editors). Gordon and Breach, N.Y., London etc. 1990, 269-463.

[42] ———, *The problem of harmonic analysis on the infinite-dimensional unitary group*, J. Funct. Anal. **205** (2003), 464–524; arXiv:math/0109193.

[43] ———, *Probability measures on dual objects to compact symmetric spaces, and hypergeometric identities*, Funkts. Analiz i Prilozh. **37** (2003), no. 4, 49–73 (Russian); English translation in Functional Analysis and its Applications **37** (2003), 281–301.

[44] ———, *Random permutations and related topics*, Chapter 25 in The Oxford Handbook on Random Matrix Theory. Gernot Akemann, Jinho Baik, and Philippe Di Francesco, eds. Oxford University Press, 2011, 510–533; arXiv:1104.1266.

[45] ———, *Projections of orbital measures, Gelfand–Tsetlin polytopes, and splines*, Journal of Lie Theory **23** (2013), no 4, 1011-1022; arXiv:1302.7116.

[46] ———, *Markov dynamics on the dual object to the infinite-dimensional unitary group*, To appear in the Proceedings of the St. Petersburg Summer School "Probability and Statistical Physics" (June 2012); arXiv:1310.6155.

[47] ———, paper in preparation.

[48] Olshanski, G. and Osinenko, A., *Multivariate Jacobi polynomials and the Selberg integral*, Functional Analysis and its Applications **46** (2012), No. 4, pp. 31–50 (Russian), pp. 262–278 (English translation).

[49] Petrov, L., *The Boundary of the Gelfand-Tsetlin Graph: New Proof of Borodin–Olshanski's Formula, and its q-Analogue*, Moscow Mathematical Journal **14** (2014), 121–160; arXiv:1208.3443.

[50] Sergeev, A. N. and Veselov, A. P., *Jacobi-Trudi formula for generalised Schur polynomials*, arXiv:0905.2557.

[51] Thoma, E., *Die unzerlegbaren, positive–definiten Klassenfunktionen der abzählbar unendlichen, symmetrischen Gruppe*, Math. Zeitschr. **85** (1964), 40–61.

[52] Vershik, A. M. and Kerov, S. V., *Asymptotic theory of characters of the symmetric group*, Funct. Anal. Appl. **15** (1981), no. 4, 246–255.

[53] ———, *Characters and factor representations of the infinite unitary group*, Doklady AN SSSR **267** (1982), no. 2, 272–276 (Russian); English translation: Soviet Math. Doklady **26** (1982), 570–574.

[54] Voiculescu, D., *Représentations factorielles de type* II_1 *de* $U(\infty)$, J. Math. Pures et Appl. **55** (1976), 1–20.

[55] Weyl, H., *The classical groups. Their invariants and representations*, Princeton Univ. Press, 1939; 1997 (fifth edition).

[56] Zelevinsky, A. V., *Representations of finite classical groups. A Hopf algebra approach*, Lecture Notes in Mathematics, **869**, Springer-Verlag, Berlin-New York, 1981.

[57] Zhelobenko, D. P., *Compact Lie groups and their representations*, Nauka, Moscow, 1970 (Russian); English translation: Transl. Math. Monographs 40, Amer. Math. Soc., Providence, RI, 1973.

Institute for Information Transmission Problems, Bolshoy Karetny 19, Moscow 127994, Russia; National Research University Higher School of Economics, Myasnitskaya 20, Moscow 101000, Russia
E-mail: olsh2007@gmail.com

Geometric intersection patterns and the theory of topological graphs

János Pach

Abstract. The *intersection graph* of a set system $\mathcal{S}$ is a graph on the vertex set $\mathcal{S}$, in which two vertices are connected by an edge if and only if the corresponding sets have nonempty intersection. It was shown by Tietze (1905) that every finite graph is the intersection graph of 3-dimensional convex polytopes. The analogous statement is false in any fixed dimension if the polytopes are allowed to have only a bounded number of faces or are replaced by simple geometric objects that can be described in terms of a bounded number of real parameters. Intersection graphs of various classes of geometric objects, even in the plane, have interesting structural and extremal properties.

We survey problems and results on geometric intersection graphs and, more generally, intersection patterns. Many of the questions discussed were originally raised by Berge, Erdős, Grünbaum, Hadwiger, Turán, and others in the context of classical topology, graph theory, and combinatorics (related, e.g., to Helly's theorem, Ramsey theory, perfect graphs). The rapid development of computational geometry and graph drawing algorithms in the last couple of decades gave further impetus to research in this field. A *topological graph* is a graph drawn in the plane so that its vertices are represented by points and its edges by possibly intersecting simple continuous curves connecting the corresponding point pairs. We give applications of the results concerning intersection patterns in the theory of topological graphs.

Mathematics Subject Classification (2010). Primary 05C35; Secondary 05C62, 52C10.

Keywords. intersection graph, topological graph, geometric graph, Ramsey theory, semialgebraic set, separator, partial order.

1. From topological graphs to intersection graphs

A *topological graph* is a graph G drawn in the *plane* with possibly intersecting curvilinear edges. More precisely, the *vertices* of G are points in the plane and the *edges* are simple continuous curves connecting the corresponding point pairs and not passing through any other point representing a vertex. These curves are allowed to cross, but we assume for simplicity that any two intersect only in a finite number of points, no two are tangent to each other, and no three share an interior point. In the special case when the edges are straight-line segments, G is called a *geometric graph*. In notation and terminology, we do not distinguish between the vertices (edges) of a topological graph and the vertices (edges) of the underlying abstract graph.

In the past few decades, the theory of topological and geometric graphs has become a fast growing separate field of combinatorial geometry with interesting applications in graph

Proceedings of the International Congress of Mathematicians, Seoul, 2014

drawing, in combinatorial and computational geometry, in additive number theory, and elsewhere. See, e.g., [5, 29, 75, 109, 114]. Many related contributions can be found in the proceedings of the annual Graph Drawing symposia, published in Springer's Lecture Notes series in Computer Science and in two collections of papers [87, 88]. For surveys, see Chapter 14 in [89], Chapter 10 in [61], and Chapters 1 and 3 in [37].

In this section, we would like to illustrate by an example how questions about topological graphs lead to the study of intersection graphs of geometric objects.

Definition 1.1. Two edges, e and f, of a topological graph are said to *cross* if they share an interior point at which e passes from one side of f to the other side. A topological graph is *simple* if any pair of its edges have at most one point in common, which is either an endpoint or an interior point at which they cross.

A topological graph is called *k-quasiplanar* for some integer $k \geq 2$ if no k of its edges are pairwise crossing.

Using this terminology, a planar graph is 2-quasiplanar.

Conjecture 1.2. *For any fixed $k \geq 2$, the number of edges of every k-quasiplanar topological graph with n vertices is $O(n)$.*

For $k = 2$, this follows from Euler's polyhedral formula. For $k = 3$, for *simple* topological graphs, Conjecture 1.2 was proved in [4]. Without the simplicity condition, the statement was first proved in [92]. The best known upper bound of roughly $8n$ was established by Ackerman and Tardos [3]. For $k = 4$, the conjecture has been verified by Ackerman [1].

For larger values of k, Conjecture 1.2 is still open. The upper bound $n(\log n)^{O(k)}$ for the number of edges of a simple k-quasiplanar topological graph was first proved in [93], and then for all k-quasiplanar topological graphs in [92]. This was further improved to $n(\log n)^{O(\log k)}$ by Fox and Pach [44]. For k-quasiplanar *geometric* graphs and, more generally, for simple topological graphs whose edges are represented by x-monotone arcs (that is, curves in the plane such that every vertical line intersects them in at most one point), Valtr [120, 121] showed that the number of edges cannot exceed $c_k n \log n$. Extending Valtr's ideas, Fox, Pach, and Suk [47] (see also [104]) proved the following.

Theorem 1.3 ([47]). *The number of edges of every k-quasiplanar topological graph of n vertices with all edges represented by x-monotone arcs is at most $2^{ck^6} n \log n$, for a suitable absolute constant c.*

Using similar ideas, Suk and Walczak [113] established another generalization of Valtr's result: the number of edges of any *simple* k-quasiplanar topological graph with n vertices is also $O_k(n \log n)$.

For *convex geometric graphs*, that is, for geometric graphs whose vertices form a convex n-gon, the Conjecture 1.2 was proved by Capoyleas and Pach [23].

Theorem 1.4 ([23]). *The maximum number of edges that a k-quasiplanar convex geometric graph with n vertices can have is $2(k-1)n - \binom{2k-1}{2}$, provided that $n \geq 2k-1$.*

The *intersection graph* of a set system $\mathcal{S}$ is a graph on the vertex set $\mathcal{S}$, in which two vertices are connected by an edge if and only if the corresponding sets have nonempty intersection.

A natural attempt to prove Conjecture 1.2, at least for geometric graphs, is the following. Let K_k denote a clique (complete graph) on k vertices.

Problem 1.5. *Given two integers $k, m > 2$, determine the smallest number $\alpha = \alpha_k(m)$ with the property that the intersection graph of any system of m segments in the plane which contains no K_k as a subgraph has at least α independent vertices. The same problem can be raised for intersection graphs of continuous curves.*

Assume for a moment that for some k there exists $\varepsilon_k > 0$ such that $\alpha_k(m) > \varepsilon_k m$. This would immediately imply Conjecture 1.2 for geometric graphs. To see this, let G be a k-quasiplanar geometric graph with vertex set $V(G)$ and edge set $E(G)$. By definition, the intersection graph of the open segments representing the edges of G contains no K_k as a subgraph. By our assumption, G has an independent set of size at least $\varepsilon_k|E(G)|$. The corresponding segments induce a planar subgraph of G. Therefore, we obtain $\varepsilon_k|E(G)| \leq 3n - 6$, implying that $|E(G)| < (3/\varepsilon_k)|V(G)|$, as required.

However, generalizing a construction of Pawlik *et al.* [100], Walczak [122] proved that $\alpha_3(m) = O(m/\log\log m)$. Hence, the above attempt to verify Conjecture 1.2 fails. The best known lower bound on $\alpha_3(m)$ is m over a polynomial in $\log m$ (see [44]). On the other hand, every K_k-free intersection graph of m unit segments in the plane has an independent set of size at least $\varepsilon_k m$, for a suitable constant $\varepsilon_k > 0$. Moreover, Suk [111] proved that for a fixed k, the chromatic numbers of these graphs are bounded by an absolute constant. In [113], a similar statement was proved for intersection graphs of continuous curves, each intersecting the x-axis in precisely one point.

Due to space limitations and personal preferences, many classical topics concerning geometric intersection patterns and graph representations will be suppressed or not mentioned at all in this survey. These include Helly-type results [123], geometric transversal theory [27], approximate embeddings of graphs into normed spaces [83], orthogonal and other geometric graph representations [78], [80], epsilon-nets and VC-dimension [107].

2. Forbidden subgraphs of intersection graphs

In combinatorics and computer science, several natural classes of geometric intersection graphs have been considered. In one dimension, the most frequently studied objects are *interval graphs*: intersection graphs of intervals. They serve as simple examples of *perfect graphs*, that is, graphs in which the chromatic number of every induced subgraph is the same as its *clique number* (the size of the largest clique). We know good characterizations of interval graphs [57, 58] in terms of forbidden subgraphs, and simple linear time algorithms for their recognition [19].

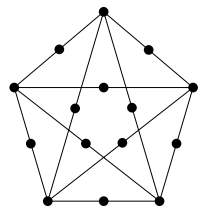
There are various natural generalizations of interval graphs in the plane: intersection graphs of (1) segments, (2) convex sets, (3) arcwise connected sets, etc. The first class is contained in the second, the second class in the third. It is easy to verify that class (3) coincides with the class of string graphs.

Definition 2.1. An intersection graph of simple continuous curves ("strings") in the plane is called a *string graph*.

The *rank* of a string graph G is the smallest integer r such that the vertices of G can be represented by continuous curves in the plane, any two of which intersect in at most r points, so that two vertices of G are connected by an edge if and only if the corresponding arcs intersect.

The investigation of string graphs was initiated by Benzer [16] and Sinden [108], in connection with genetic structures and printed electrical circuits.

Sinden [108] showed that the graph of *fifteen* vertices depicted below is not a string graph, therefore, it does not belong to any of the classes (1)-(3). Ten years later Ehrlich, Even, and Tarjan [28] constructed a string graph which is not a segment intersection graph, that is, it belongs to class (3), but not to class (1).



Thus, a string graph cannot contain this 15-vertex graph as an induced subgraph. However, we cannot hope that string graphs have a good characterization in terms of forbidden subgraphs, unless P=NP: it was shown by Kratochvíl [69] that recognizing string graphs is NP-hard, and by Schaefer, Sedgwick, and Štefankovič [106] that it belongs to NP. The problem of recognizing whether a graph is an intersection graph of segments is also known to be NP-hard (in fact, equivalent to the existential theory of reals [70, 105]). Even for relatively simple graphs, it may be a formidable task to decide whether they allow such a representation by segments. For example, it was a longstanding conjecture that every finite planar graph is an intersection graph of segments. It was finally verified by Chalopin and Gonçalves in 2009 [24].

Definition 2.2. A graph property P is called *hereditary* if every induced subgraph of a graph with property P also has property P. The property that G is a string graph obviously satisfies this condition.

The total number of graphs on n labeled vertices is $2^{\binom{n}{2}}$. Since most of them contain an induced subgraph isomorphic to the *fifteen*-vertex graph depicted above, most graphs are not string graphs. Using the extremal theory of graphs with some hereditary property, developed in [18, 32, 103], Pach and Tóth [97] established the following more precise asymptotic results.

Theorem 2.3 ([97]). *The number of string graphs on n labeled vertices is $2^{(\frac{3}{4}+o(1))\binom{n}{2}}$.*

Theorem 2.4 ([97]). *For any fixed positive k, the number of string graphs of rank k on n labeled vertices is $2^{o(n^2)}$.*

Every graph which is an intersection graph of segments is a string graph of rank 1. For intersection graphs of segments, we have a much better result, which can be deduced using a theorem of Oleĭnik and Petrovsky [101], Milnor [86], and Thom [117] from real algebraic geometry (see also [15]). The number of n-vertex intersection graphs of segments is $2^{O(n \log n)}$, and the order of magnitude of the exponent is correct [94]. Fox (personal communication) has recently proved the asymptotically tight bound $2^{(4+o(1))n \log n}$. The best known upper bound, $2^{O(n^{3/2} \log n)}$, for the number of n-vertex string graphs of rank 1 is due to Kynčl [73]. The structure of a "typical" string graph was studied in [62].

The above results can be applied to estimate the number of combinatorially different ways a complete graph on n vertices can be drawn in the plane so that any pair of its edges cross at most r times, where r is a fixed positive integer [97], [73].

3. Ramsey-type properties of intersection graphs

By the quantitative form of Ramsey's theorem, established by Erdős and Szekeres [35], every graph of n vertices has a clique or an independent set of size at least $\frac{1}{2}\log_2 n$. In a seminal paper written in 1989, Erdős and Hajnal [33] showed that the family of all graphs that do not contain a fixed forbidden graph G as an induced subgraph, have much stronger Ramsey-type properties than the family of all graphs. More precisely, they proved the following.

Theorem 3.1 ([33]). *For any graph G, there exists a constant $c = c(G) > 0$ such that every graph of n vertices that does not contain G as an induced subgraph has a clique or an independent set of size at least $e^{c\sqrt{\log n}}$.*

They raised the question whether one can always find a complete or empty induced subgraph of size n^c. This remains one of the most challenging open problems in Ramsey theory.

A complete bipartite graph with $\lceil n/2 \rceil$ vertices in one class and $\lfloor n/2 \rfloor$ vertices in the other is called a *bi-clique of size n*. Erdős, Hajnal, and Pach [34] proved a bipartite variant.

Theorem 3.2 ([34]). *For any graph G, there is a constant $c = c(G) > 0$ such that every graph on n vertices that does not contain G as an induced subgraph has a bi-clique of size n^c or the complement of such a bi-clique.*

See [51] for a strengthening of this result.

Obviously, the last two theorems remain true for all hereditary families of graphs, that is, for any family other than the family of all finite graphs that is closed under taking induced subgraphs. The family of string graphs (intersection graphs of continuous curves or arcwise connected sets in the plane) and, hence, the families of all graphs that can be obtained as intersection graphs of segments, convex sets, etc. belong to this category.

In [40], we introduced the following terminology.

Definition 3.3. A family $\mathcal{F}$ of graphs has the

1. *(Weak) Erdős-Hajnal property* if there is a constant $c(\mathcal{F}) > 0$ such that every graph in $\mathcal{F}$ on n vertices contains a clique or an independent set of size $n^{c(\mathcal{F})}$;
2. *Strong Erdős-Hajnal property* if there is a constant $b(\mathcal{F}) > 0$ such that for every graph G in $\mathcal{F}$ on n vertices contains a bi-clique of size $b(\mathcal{F})n$ or the complement of such a bi-clique.

It was shown in [6] that if a hereditary family of graphs has the strong Erdős-Hajnal property, then it also has the Erdős-Hajnal property. The converse is false, as is shown, e.g., by the family of triangle-free graphs. The first nontrivial result showing that a geometric intersection graph has the Erdős-Hajnal property was found by Larman *et al.* [74].

Theorem 3.4 ([74]). *The intersection graph of n convex sets in the plane has a clique or an independent set of size at least $n^{1/5}$.*

It is enough to assume here that every set of the family is *vertically convex*, that is, a connected set with the property that every vertical line meeting it intersects it in an interval or in a point. It is an interesting open problem to improve the exponent $1/5$ in the theorem. The best known upper bound, due to Kynčl [72], is $\log 8/\log 169 \approx .405$ (cf. [63]), so there is plenty of room for improvement.

The family of intersection graphs of convex sets in the plane also has the *strong* Erdős-Hajnal property [50]. However, the family of intersection graphs of vertically convex sets does not [96]. By definition, any *x-monotone* curve, that is, any continuous curve in the plane such that every vertical line intersects it in at most one point, is vertically convex.

Theorem 3.5 ([96]). *For every n, there is an n-member family of x-monotone curves in the plane such that neither their intersection graph, nor its complement contains a bi-clique of size at least $cn/\log n$. Here c is an absolute constant.*

In the other direction, it is known that every string graph with n vertices or its complement contains a bi-clique of size at least $c'n/\log n$ (cf. [45]).

If we put an upper bound r on the number of times two curves are allowed to meet, then the corresponding intersection graphs, string graphs of rank r (see Definition 2.1) behave much nicer.

Theorem 3.6 ([49]). *The family of string graphs of rank r has the strong Erdős-Hajnal property.*

One of the most challenging unsolved problems in this area is to decide whether the family of *all* string graphs has the (weak) Erdős-Hajnal property. The best known result in this direction was established in [46]: Every string graph with n vertices has a clique or an independent set of size at least $n^{c/\log\log n}$.

4. Intersection graphs of semialgebraic sets

According to Tietze's theorem [118] cited in the abstract, every finite graph can be obtained as the intersection graph of 3-dimensional convex bodies. This may suggest that there is no hope to generalize the results in the previous section to higher dimensions. Actually, this is not the case. The proof method of Pach-Solymosi [94], where it was first shown that the family of intersection graphs of segments in the plane has the strong Erdős-Hajnal property, can be extended as follows.

Definition 4.1 ([15]). A *semialgebraic set* S in $\mathbb{R}^d$ is the locus of all points that satisfy a given finite Boolean combination of at most d polynomial equations and inequalities of degree at most d in the d coordinates. (Without loss of generality, these three parameters are bounded by the same integer d.) The *description complexity* of S is the smallest integer d for which S has such a representation.

Every element S of a family $\mathcal{F}$ of semialgebraic sets of constant description complexity d can be represented by a point S^* of a d^*-dimensional Euclidean space (in which the coordinates are, say, the coefficients of the monomials in the polynomials that define S). A graph (binary relation) $R \subset \mathcal{F} \times \mathcal{F}$ is *semialgebraic* if the corresponding set $\{(S^*, T^*) \in \mathbb{R}^{2d^*} \mid S, T \in \mathcal{F}, (S,T) \in R\}$ is semialgebraic. *Semialgebraic hypergraphs* (*relations* of h variables, h-ary relations) can be defined analogously.

Theorem 4.2 ([6]). *For any d, the family of all graphs that are associated with a semialgebraic binary relation of description complexity at most d has the strong Erdős-Hajnal property.*

The relation that two semialgebraic sets, $S, T \in \mathcal{F}$, with description complexity d have nonempty intersection is semialgebraic. Thus, we have the following.

Corollary 4.3 ([6]). *Any family of intersection graphs of (real) semialgebraic sets of constant description complexity has the strong (and, therefore, the weak) Erdős-Hajnal property.*

Basu [14] extended this result for a broader class of algebraically defined sets (o-minimal sets).

An n-vertex graph is called *t-Ramsey* if it contains no clique and no independent set of size at least t. A probabilistic construction of Erdős [31] shows that almost all n-vertex graphs are $2\log_2 n$-Ramsey, but it is a formidable task to find comparably good efficient constructions. The best known polynomial time deterministic algorithm, due to Barak *et al.* [11], produces only $2^{(log n)^{o(1)}}$-Ramsey graphs. The previous record was held by Frankl and Wilson [52]. Theorem 4.2 above shows that no $n^{o(1)}$-Ramsey graphs can be defined using semialgebraic relations of constant description complexity. This settles a conjecture of Babai [10].

Fox, Gromov, Lafforgue, Naor, and Pach [39] proved the following far-reaching generalization of Theorem 4.2.

Theorem 4.4 ([39]). *Let $\alpha > 0$, let $\mathcal{F}_1, \ldots, \mathcal{F}_h$ be finite families of semialgebraic sets of constant description complexity, and let R be a fixed semialgebraic h-ary relation on $\mathcal{F}_1 \times \cdots \times \mathcal{F}_h$ such that the number of h-tuples that are related (resp. unrelated) with respect to R is at least $\alpha \prod_{i=1}^{h} |\mathcal{F}_i|$. Then there exists a constant $c' > 0$, which depends on α, h and on the maximum description complexity d of the sets in $\mathcal{F}_i$ $(1 \le i \le h)$ and R, and there exist subfamilies $\mathcal{F}_i' \subseteq \mathcal{F}_i$ with $|\mathcal{F}_i'| \ge c'|\mathcal{F}_i|$ $(1 \le i \le h)$ such that $\mathcal{F}_1' \times \cdots \times \mathcal{F}_h' \subseteq R$ (resp. $(\mathcal{F}_1' \times \cdots \times \mathcal{F}_h') \cap R = \emptyset$). Moreover, each subset $\mathcal{F}_i'$ consists of exactly those elements of $\mathcal{F}_i$ that satisfy a certain semialgebraic relation of constant description complexity.*

Apart from the fact that the last statement also handles semialgebraic hypergraphs (h-ary relations), it also strengthens Theorem 4.2 in another direction. It is not just a Ramsey-type theorem, which guarantees that at least one of two or several possibilities will occur. It is a so-called "density theorem," which tells us that if sufficiently many h-tuples are related by the relation R (that is, the h-uniform semialgebraic hypergraph R has sufficiently many hyperedges), then there are h large subsets $\mathcal{F}_i' \subseteq \mathcal{F}_i$ $(1 \le i \le h)$ such that no matter how we pick an element from each, the resulting h-tuple is related (is a hyperedge of R). The constant c' in Theorem 4.4 can be taken to a polynomial in α (see [48]).

By repeated application of this statement, one can obtain an even stronger Szemerédi-type partition theorem. An *equipartition* of a finite set P is a partition $P = P_1 \cup \ldots \cup P_k$ into almost equal parts. That is, $|P_i| = \lfloor |P|/k \rfloor$ or $\lceil |P|/k \rceil$ for every i.

Theorem 4.5 ([39]). *For any h, d and for any $\varepsilon > 0$, there exists $K = K(\varepsilon, h, d)$ satisfying the following condition. For any $k \ge K$, for any semialgebraic relation R on h-tuples of points in a Euclidean space $\mathbb{R}^d$ with description complexity at most d, every finite set $P \subseteq \mathbb{R}^d$ has an equipartition $P = P_1 \cup \ldots \cup P_k$ such that all but at most an ε-fraction of the h-tuples $(P_{i_1}, \ldots, P_{i_h})$ have the property that either all r-tuples of points with one element in each P_{i_j} are related with respect to R or none of them are.*

It was shown in [48] that the constant K in Theorem 4.5 can be bounded from above by a polynomial of $1/\varepsilon$.

The investigation of semialgebraic versions of Ramsey's theorem for h-ary relations was initiated in [25]. Let $N_h^d(n)$ be the smallest integer N such that for any semialgebraic relation R on h-tuples of N points in $\mathbb{R}^d$ with description complexity at most d, there is a *homogeneous* subset of size n, that is, a subset with the property that either all of its h-tuples belong to R or none of them does. It was shown that the function $N_h^d(n)$ grows in n as a tower of height $h-1$, and that in some sense this result is optimal. This is one exponential better than the behavior of the general Ramsey function for arbitrary h-ary relations.

For some related results and geometric applications, see [12, 13, 20, 30, 39, 112].

5. Intersection graphs and partially ordered sets

Given a partially ordered set $(P, <)$, its *incomparability graph* is the graph with vertex set P, in which two elements are adjacent if and only if they are incomparable. Incomparability graphs are fairly well understood. In 1950, Dilworth [26] proved that every incomparability graph is a perfect graph, so the chromatic number of an incomparability graph is equal to its clique number. Gallai [57] gave a characterization of incomparability graphs in terms of minimal forbidden induced subgraphs, and there exist polynomial time algorithms to recognize them [59].

There is a curious relation between incomparability graphs and string graphs (Definition 2.1), which was first observed by Golumbic, Rotem, and Urrutia [60] and, independently, by Lovász [77].

Theorem 5.1 ([60, 77]). *Every incomparability graph is a string graph.*

The converse is obviously not true. For example, a cycle of length *five* is a string graph, but it is not perfect, therefore, it cannot be an incomparability graph. Kleitman and Rothschild [65] showed that the number of incomparability graphs on n vertices is only $2^{(1/2+o(1))\binom{n}{2}}$, which is much smaller then the number of string graphs, asymptotically given in Theorem 2.3.

Nevertheless, it was shown by Fox and Pach [45] that most string graphs contain huge subgraphs that are incomparability graphs. The geometric conditions somehow seem to enforce a partial order on the curves.

Theorem 5.2. *For every $\varepsilon > 0$ there exists $\delta > 0$ with the property that if $\mathcal{F}$ is a family of curves whose string graph has at least $\varepsilon|\mathcal{F}|^2$ edges, then one can select a subcurve γ' of each $\gamma \in \mathcal{F}$ such that the string graph of the family $\{\gamma' : \gamma \in \mathcal{F}\}$ has at least $\delta|\mathcal{F}|^2$ edges and is an incomparability graph.*

This implies that every dense string graph contains a dense *spanning* subgraph (i.e., a dense subgraph on the same vertex set) which is an incomparability graph. However, it is not true that every dense string graph contains a dense *induced* subgraph with a linear number of vertices that is an incomparability graph. Indeed, since every incomparability graph is perfect, this would imply that every string graph has a clique or an independent set of size at least constant times $\sqrt{n}$. This is certainly false, e.g., for the construction of Kynčl, mentioned after Theorem 3.4.

Fox [38] proved that incomparability graphs "almost" have the strong Erdős-Hajnal property.

Theorem 5.3 ([38]). *If n is large enough, the incomparability graph of every n-element partially ordered set, or its complement, the comparability graph, has a bi-clique of size at least $\frac{n}{4\log_2 n}$. This bound is tight up to a constant factor.*

The second part of this statement, combined with Theorem 5.1, immediately implies that the family of string graphs does not have the strong Erdős-Hajnal property (which was Theorem 3.5).

In [42], Theorem 5.3 was generalized to several partial orders. Note that the proof of Theorem 3.4 is based on the fact that on any family of (vertically) convex sets in the plane, we can define *four* partial orders so that two sets have nonempty intersection if and only if they are incomparable by all of them. Using the generalized version of Theorem 5.3, we obtain that the intersection graphs of (vertically) convex sets in the plane also "almost" have the strong Erdős-Hajnal property. As was mentioned in Section 3, for convex sets a stronger statement is true (which does not hold under the weaker assumption of vertical convexity).

Theorem 5.4 ([50]). *The family of intersection graphs of finitely many convex sets in the plane has the strong Erdős-Hajnal property.*

The *dimension* of a partially ordered set $(P, >)$ is the minimum number of linear extensions of the relation ">" such that their intersection is ">". For the proof of Theorem 5.4, one has to consider a new type of extremal problem for incomparability graphs: What is the maximum number of edges that an n-vertex incomparability graph of a partial order of dimension d can have if it does not contain, say, a complete bipartite subgraph $K_{r,r}$, for a fixed r? The same question can be asked about comparability graphs and also for the case where the condition on the dimension is dropped.

In the same paper, a stronger form of Theorem 5.3 was proved for dense graphs.

Theorem 5.5 ([50]). *For every $\varepsilon > 0$, there exists $\delta > 0$ such that every incomparability graph with n vertices and at least εn^2 edges contains a bi-clique of size $\delta n/\log n$.*

Combining this result with Theorem 5.1, we obtain

Corollary 5.6. *For every $\varepsilon > 0$, there exists $\delta > 0$ such that every string graph with n vertices and at least εn^2 edges contains a bi-clique of size $\delta n/\log n$.*

The formulation of Theorem 5.3 may suggest a certain kind of symmetry between incomparability and comparability graphs. However, Theorem 5.1 has no analogue for comparability graphs. The following strengthening of Theorem 5.3 is also slightly asymmetric.

Theorem 5.7 ([50]). *There is constant $c > 0$ such that the incomparability graph of every n-element partially ordered set has a bi-clique of size at least $cn/\log n$, or its complement, the comparability graph, has a bi-clique of size at least cn.*

It was conjectured in [40] that every perfect graph on n vertices or its complement contains a bi-clique of size at least $n^{1-o(1)}$.

6. Intersection graphs and planar separators

Given a family of continuous curves (strings) in the plane, introducing a vertex at each intersection point and each endpoint of the curves, we obtain a planar graph. Under some fairly natural conditions, there are few strings that connect far-away parts of this planar graph. In such cases, there is a good chance that we can use the *Lipton-Tarjan separator theorem* for planar graphs [76].

A *separator* for a graph $G = (V, E)$ is a subset $V_0 \subset V$ such that there is a partition $V = V_0 \cup V_1 \cup V_2$ with $|V_1|, |V_2| \leq \frac{2}{3}|V|$ and no vertex in V_1 is adjacent to any vertex in V_2. The Lipton-Tarjan separator theorem states that every planar graph with n vertices has a separator of size $O(\sqrt{n})$. By a classical theorem of Koebe [66], every planar graph can be represented as the intersection graph of closed disks in the plane with disjoint interiors. Miller, Teng, Thurston, and Vavasis [85] found a generalization of the Lipton-Tarjan separator theorem to higher dimensions. They proved that the intersection graph of any family of n balls in $\mathbb{R}^d$ such that no k of them have a point in common has a separator of size $O(dk^{1/d}n^{1-1/d})$.

Fox and Pach [41] established the following common generalization of the separator theorems of Lipton and Tarjan and of Miller *et al.* in the plane.

Theorem 6.1 ([41]). *If $\mathcal{F}$ is a finite family of Jordan regions with a total of m boundary crossings, then the intersection graph of $\mathcal{F}$ has a separator of size $O(\sqrt{m})$.*

Corollary 6.2 ([41]). *If $\mathcal{F}$ is a finite family of curves in the plane with a total of m crossings, then the intersection graph (string graph) of $\mathcal{F}$ has a separator of size $O(\sqrt{m})$.*

Using Theorem 6.1 and Theorem 1.4, one can deduce the following.

Theorem 6.3 ([41]). *Every K_k-free intersection graph of convex bodies in the plane with m edges has a separator of size $O(\sqrt{km})$.*

Notice that in this statement, the size of the separator is bounded in terms of the number of edges of the intersection graph, rather than the number of vertices. Nevertheless, since planar graphs are K_5-free and (by Koebe's theorem) can be obtained as intersection graphs of convex bodies, Theorem 6.3 also implies the Lipton-Tarjan separator theorem.

Fox and Pach [43] made the following conjecture, much stronger than Corollary 6.2.

Conjecture 6.4 ([43]). *Every string graph with m edges has a separator of size $O(\sqrt{m})$.*

In [43], a weaker bound, $O(m^{3/4}\sqrt{\log m})$, was established. This bound was used to deduce the following interesting property of string graphs. Let $K_{k,k}$ denote the complete bipartite graph with k vertices in each of its classes (that is, a bi-clique of size $2k$).

Theorem 6.5 ([43]). *For any positive integer k, there is a constant $c(k)$ such that every $K_{k,k}$-free string graph with n vertices has at most $c(k)n$ edges.*

This is in sharp contrast with the general behavior of graphs. According to the Kővári-Sós-Turán theorem [68], for a fixed k, every $K_{k,k}$-free graph with n vertices has at most $O(n^{2-1/k})$ edges. For $k > 2$, this bound is not known to be optimal, but the right exponent is definitely at least $2 - 2/k > 1$ (see, e.g., [17]). It is a rich and active subfield of extremal graph theory to estimate the maximum number of edges of a B-free graph of n vertices, for a given bipartite graph B. Theorem 6.5 shows that for string graphs there is no such theory: no matter what B is, the maximum is $O(n)$.

Matoušek [84] came close to proving Conjecture 6.4. He adapted some powerful techniques developed by Feige, Hajiaghayi, and Lee [36], who used the framework of multicommodity flows to design efficient approximation algorithms for finding small separators. See also [67].

Theorem 6.6 ([84]). *Every string graph with m edges has a separator of size at most $O(\sqrt{m}\log m)$.*

In [46], the last theorem was utilized to deduce that Theorem 6.5 is true with $c(k) = k(\log k)^{O(1)}$, which is not far from being optimal. It is conjectured that the best possible value of $c(k)$ for which the theorem still holds satisfies $c(k) = O(k \log k)$.

7. The theory of topological graphs

It was probably Erdős who first suggested in the 1960s that some of the basic questions in extremal graph theory have natural analogues for geometric or topological graphs. For instance, what is the maximum number of edges that a geometric graph of n vertices can have without containing a fixed "forbidden" configuration, that is, a set of edges such that their intersection pattern is specified. The first such result, in which the forbidden configuration consisted of 2 disjoint edges (that cannot have any endpoints or internal points in common) was published by Avital and by Erdős's close friend, Hanani [9]. The answer is n. Thirteen years later, in his master's thesis [71], Kupitz started to explore these questions systematically. Alon and Erdős [7] proved that every geometric graph with no 3 disjoint edges has $O(n)$ edges. The first general bound was established in [99] and uses partial orders.

Theorem 7.1 ([99]). *For any integer $k \geq 2$, the maximum number of edges of a geometric graph with n vertices that contains no k disjoint edges is $O_k(n)$.*

The best known value of the constant hidden in the O_k-notation is $O(k^2)$ (see [119]). It is perfectly possible that this bound can be improved to $O(k)$, which would be best possible.

It is conjectured that Theorem 7.1 remains true for simple topological graphs, i.e., for topological graphs in which every pair of edges intersect in at most one point (Definition 1.1). For the case $k = 2$, Conway made the following stronger conjecture, which has become known as the "thrackle conjecture".

Conjecture 7.2 ([124]). *Every simple topological graph with $n \geq 3$ vertices that contains no 2 disjoint edges has at most n edges.*

It is known that every such graph has a linear number of edges in n (see [21, 53, 79]). The thrackle conjecture has been verified for simple topological graphs with x-monotone edges ([95], cf. Theorem 3.5) and in the case where all vertices lie on a circle and all edges in its interior [22].

We do not know whether the analogue of Theorem 7.1 is true for simple topological graphs, when $k \geq 3$. All we know is that, according to [98], the maximum number of edges of a simple topological graph with n vertices that contains no k disjoint edges is $n(\log n)^{O(k)}$. In the most optimistic scenario, this bound could be improved to $O(kn)$. Suppose that this is the case. This would imply that a complete simple topological graph with n vertices (and $\binom{n}{2}$ edges) must have at least cn disjoint edges, for a suitable constant $c > 0$. Suk [110] proved a weaker bound.

Theorem 7.3 ([110]). *Every complete simple topological graph with n vertices must have at least $cn^{1/3}$ disjoint edges, for a suitable constant $c > 0$.*

An alternative proof of this theorem was found by Fulek and Ruiz-Vargas [54]. Ruiz-Vargas has recently announced the improved bound $cn^{1/2-\varepsilon}$, for every $\varepsilon > 0$. Both proofs break down if we want to extend Theorem 7.3 to all *dense* simple topological graphs, that is, to graphs with at least δn^2 edges for some $\delta > 0$. We cannot generalize this statement even for complete *bipartite* simple topological graphs.

In Section 1, we considered the "dual" problem, where the forbidden configuration consists of k pairwise *crossing* edges. Recall that topological graphs with no k pairwise crossing edges are called *k-quasiplanar* (see Definition 1.1). What is the maximum number of edges that a k-quasiplanar topological graph of n vertices can have? The conjectured answer is $O_k(n)$ (or perhaps even $O(kn)$; cf. Conjecture 1.2). As was mentioned in Section 1, this is known to be true only for $k \leq 4$. Presently, the best upper bound is $n(\log n)^{O(\log k)}$.

If the stronger conjecture was true, i.e., every k-quasiplanar graph of n vertices had at most $O(kn)$ edges, it would follow that every complete topological graph of n vertices has at least cn pairwise crossing edges, for a suitable constant $c > 0$. For geometric graphs, Aronov *et al.* [8] established a weaker statement, dual to Theorem 7.3: Every complete geometric graph with n vertices must have at least $cn^{1/2}$ pairwise crossing edges, for a suitable constant $c > 0$. A similar statement holds for all reasonably dense topological graphs, in which any pair of edges intersect at most a bounded number of times.

Theorem 7.4 ([44]). *For every $\varepsilon > 0$ and for every integer $t > 0$, there exists $\delta = \delta(\varepsilon, t) > 0$ with the following property. Every topological graph with n vertices, in which no two edges intersect in more than t points, has at least n^{δ} pairwise crossing edges.*

It follows from the results in [46] that if we drop the assumption in the last theorem that every pair of edges intersect in at most t points, then we can guarantee the existence of only $n^{\delta/\log\log n}$ pairwise crossing edges.

More complicated forbidden configurations have also been considered. For instance, let k be a positive integer and let G be a geometric graph with n vertices that contains no two sets of edges, $E_1, E_2 \subset E(G)$, each consisting of k pairwise crossing edges, such that every edge in E_1 is disjoint from every edge in E_2. Fulek and Suk [55] proved that then G has at most $O_k(n \log n)$ edges, and they conjectured that the correct order of magnitude is linear for every fixed k.

Let k and l be fixed positive integers. A *(k, l)-grid* in a topological graph is a pair of subsets, $E_1, E_2 \subset E(G)$, with $|E_1| = k, |E_2| = l$ such that every edge in E_1 crosses every edge in E_2. If, in addition, each E_i consists of disjoint edges, the (k, l)-grid is called *natural*. It is known that every n-vertex topological graph with no (k, l)-grid has $O_{k,l}(n)$ edges [90], [116].

Conjecture 7.5 ([2]). *For any positive integers k and l, there exists a constant $c_{k,l}$ such that every simple topological graph on n vertices with no natural (k, l)-grid has at most $c_{k,l}n$ edges.*

This conjecture would immediately imply that Theorem 7.1 generalizes to simple topological graphs. It would also imply that every simple topological graph on n vertices which contains no (k, l)-grid such that all $2(k + l)$ endpoints of its edges are distinct, has at most $O_{k,l}(n)$ edges. We cannot even verify this weaker conjecture. We can prove only the following.

Theorem 7.6 ([2]). *For every positive integer k, there is a constant c_k such that every topological graph on n vertices that contains no (k,k)-grid with distinct vertices has at most $c_k n \log^* n$ edges, where $\log^*$ denotes the iterated logarithm function.*

It was already pointed out by Klazar and Marcus [64], in a slightly different formulation, that the proof of the Marcus-Tardos theorem [81] can be easily modified to prove that Conjecture 7.5 is true for *convex* geometric graphs, that is, for geometric graphs whose vertices form the vertex set of a convex n-gon.

The above mentioned results and conjectures might suggest that for every nontrivial forbidden configuration F of a fixed size, the maximum number of edges that an F-free geometric or topological graph with n vertices can have is linear in n. However, this is not the case. It was shown in [91] that the maximum number of edges of a geometric graph with n vertices, containing no self-intersecting path of length 3, is at most $cn \log n$ for a suitable constant c, and that the order of magnitude of this bound cannot be improved. This result was extended by Tardos [115]: for every $k \geq 3$, he constructed geometric graphs with a superlinear number of edges that contain no self-intersecting path of length k. As a corollary, one can obtain a simple characterization of all abstract graphs G, for which all geometric graphs with n vertices that contain no self-intersecting subgraph isomorphic to G have $O(n)$ edges: these graphs are forests with at least two components that are not isolated vertices.

Note that there exist arbitrarily large (abstract) graphs with a superlinear number of edges that contain no cycle of a fixed length k. For example, it is well known that for $k = 4$, there are C_4-free graphs with n vertices and $(\frac{1}{2} + o(1))n^{3/2}$ edges (see [17, 56]). On the other hand, improving an argument of Pinchasi and Radoičić [102], Marcus and Tardos [82] obtained the following almost tight result.

Theorem 7.7 ([82]). *Every topological graph on n vertices that contains no self-intersecting cycle of length 4 has at most $O(n^{3/2} \log n)$ edges.*

Acknowledgements. The author is supported by OTKA grant NN-102029 under EuroGIGA projects GraDR and ComPoSe, and by Swiss National Science Foundation Grants 200020-144531 and 200021-137574.

References

[1] Ackerman, E., *On the maximum number of edges in topological graphs with no four pairwise crossing edges*, Discrete Comput. Geom. **41** (2009), 365–375.

[2] Ackerman, E., Fox, J., Pach, J., and Suk, A., *On grids in topological graphs*, in: 25th Symp. Comput. Geometry (SoCG 2009), ACM Press, New York, 2009, 403–412.

[3] Ackerman, E. and Tardos, G., *On the maximum number of edges in quasi-planar graphs*, J. Combin. Theory, Ser. A **114** (2007), 563–571.

[4] Agarwal, P. K., Aronov, B., Pach, J., Pollack, R., and Sharir, M., *Quasi-planar graphs have a linear number of edges*, Combinatorica **17** (1997), 1–9.

[5] Ajtai, M., Chvátal, V., Newborn, M., and Szemerédi, E., *Crossing free graphs*, Ann. Discrete Math. **12** (1982) 9–12.

[6] Alon, N., Pach, J., Pinchasi, R., Radoičić, R., and Sharir, M., *Crossing patterns of semi-algebraic sets*, J. Combin. Theory, Ser. A **111** (2) (2005), 310–326.

[7] Alon, N. and Erdős, P., *Disjoint edges in geometric graphs*, Discrete Comput. Geom. **4** (1989), 287–290.

[8] Aronov, B., Erdős, P., Goddard, W., Kleitman, D. J., Klugerman, M., Pach, J., and Schulman, L. J., *Crossing families*, Combinatorica **14** (1994), no. 2, 127–134.

[9] Avital, S. and Hanani, H., *Graphs, continuation*, Gilyonot Le'matematika **3** (1966), no. 2, 2–8.

[10] Babai, L., *Open problem*, in: Proc. 5th Hungar. Conf. Combin. (A. Hajnal and V. T. Sos, eds.), Keszthely, Hungary, 1976, Vol. 2, North Holland (1978), 1189.

[11] Barak, B., Rao, R., Shaltiel, R., and Wigderson, A., *2-source dispersers for $n^{o(1)}$ entropy, and Ramsey graphs beating the Frankl-Wilson construction*, Ann. of Math. (2) **176** (2012), no. 3, 1483–1543.

[12] Bárány, I., Matoušek, J., and Pór, A., *Curves in $\mathbb{R}^d$ intersecting every hyperplane at most $d+1$ times*, in: Proc. 30th Symp. Comput. Geometry (SoCG 2014), to appear.

[13] Bárány, I. and Pach, J., *Homogeneous selections from hyperplanes*, J. Combin. Theory, Ser. B **104** (2014), 81–87.

[14] Basu, S., *Combinatorial complexity in o-minimal geometry*, Proc. Lond. Math. Soc. (3) **100** (2010), no. 2, 405–428.

[15] Basu, S., Pollack, R., Roy, M.-F., *Algorithms in Real Algebraic Geometry. Second ed*, Algorithms and Computation in Mathematics **10**, Springer-Verlag, Berlin, 2006.

[16] Benzer, S., *On the topology of the genetic fine structure*, Proc. Nat. Acad. Sci. of USA **45/11** (1959), 1607–1620.

[17] Bollobás, B., *Modern Graph Theory*, Springer-Verlag New York, 1998.

[18] Bollobás, B. and Thomason, *Hereditary and monotone properties of graphs*, in: The Mathematics of Paul Erdős II, (R. L. Graham and J. Nešetřil, eds.), Algorithms Combin. **14**, Springer, Berlin, 1997, 70–78.

[19] Booth, K. S. and Lueker, G. S., *Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms*, J. Comput. System Sci. **13** (1976), 335–379.

[20] Bukh, B. and Hubard, A., *Space crossing numbers*, Combin. Probab. Comput. **21** (2012), no. 3, 358–373.

[21] Cairns, G. and Nikolayevsky, Y., *Bounds for generalized thrackles*, Discrete Comput. Geom. **23** (2000), 191–206.

[22] ———, *Outerplanar thrackles*, Graphs Combin. **28** (2012), 85–96.

[23] Capoyleas, V. and Pach, J., *A Turán-type theorem on chords of a convex polygon*, J. Combin. Theory, Ser. B **56** (1992), 9–15.

[24] Chalopin, J. and Gonçalves, D. *Every planar graph is the intersection graph of segments in the plane: extended abstract*, STOC 2009, 631–638.

[25] Conlon, D., Fox, J., Pach, J., Sudakov, B., and Suk, A., *Ramsey-type results for semi-algebraic sets*, in: Proc. 29th Symp. Comput. Geometry (SoCG 2013), ACM Press, New York, 2013, 309–318.

[26] Dilworth, R. P., *A decomposition theorem for partially ordered sets*, Annals of Math. **51**, (1950), 161–166.

[27] Eckhoff, J., *A survey of the Hadwiger-Debrunner* (p, q)*-problem*, in: Discrete and Computational Geometry, Algorithms Combin. **25**, Springer, Berlin, 2003, 347–377.

[28] Ehrlich, G., Even, S., and Tarjan, R. E., *Intersection graphs of curves in the plane*, J. Combin. Theory, Ser. B **21** (1976), no. 1, 8–20.

[29] Elekes, G., *On the number of sums and products*, Acta Arith. **81** (1997), 365–367.

[30] Eliáš, M., Matoušek, J., Roldán-Pensado, E., and Safernová, Z., *Lower bounds on geometric Ramsey functions*, in: Proc. 30th Symp. Comput. Geometry (SoCG 2014), to appear.

[31] Erdős, P., *Some remarks on the theory of graphs*, Bull. Amer. Math. Soc. **53** (1947), 292–294.

[32] Erdős, P., Frankl, P., and Rödl, V., *The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent*, Graphs and Combinatorics **2** (1986), 113–121.

[33] Erdős, P. and Hajnal, A., *Ramsey-type theorems*, Discrete Appl. Math. **25** (1989), 37–52.

[34] Erdős, P., Hajnal, A., and Pach, J., *Ramsey-type theorem for bipartite graphs*, Geombinatorics **10** (2000), 64–68.

[35] Erdős, P. and Szekeres, G., *A combinatorial problem in geometry*, Compositio Mathematica **2** (1935), 463–470.

[36] U. Feige, M. Hajiaghayi, and Lee, J. R., *Improved approximation algorithms for minimum weight vertex separators*, SIAM J. Comput. **38** (2008), no. 2, 629–657.

[37] Felsner, S., *Geometric Graphs and Arrangements*, Vieweg & Sohn, Wiesbaden, 2004.

[38] Fox, J., *A bipartite analogue of Dilworth's theorem*, Order **23** (2-3) (2006), 197–209.

[39] Fox, J., Gromov, M., Lafforgue, V., Naor, A., and Pach, J., *Overlap properties of geometric expanders*, J. Reine Angew. Math. **671** (2012), 49–83.

[40] Fox, J. and Pach, J., *Erdős-Hajnal-type results on intersection patterns of geometric objects*, in: Horizons of Combinatorics, Bolyai Soc. Math. Stud. **17**, Springer, Berlin, 2008, 79–103.

[41] ———, *Separator theorems and Turán-type results for planar intersection graphs*, Advances in Mathematics **219** (2008), 1070–1080.

[42] _____, *A bipartite analogue of Dilworth's theorem for multiple partial orders*, European J. Combin. **30** (2009), no. 8, 1846–1853.

[43] _____, *A separator theorem for string graphs and its applications*, Combin. Probab. Comput. **19** (2010), no. 3, 371–390.

[44] _____, *Coloring K_k-free intersection graphs of geometric objects in the plane*, European J. Combin. **33** (2012), 853–866.

[45] _____, *String graphs and incomparability graphs*, Adv. Math. **230** (2012), no. 3, 1381–1401.

[46] _____, *Applications of a new separator theorem for string graphs*, Combin. Probab. Comput. **23** (2014), no. 1., 66–74.

[47] Fox, J., Pach, J., and Suk, A., *The number of edges in k-quasi-planar graphs*, SIAM J. Discrete Math. **27** (2013), 550–561.

[48] _____, *Density and regularity theorems for semi-algebraic hypegraphs*, manuscript.

[49] Fox, J., Pach, J., and Tóth, C. D., *A bipartite strengthening of the crossing lemma*, J. Combin. Theory, Ser. B **100** (2010), 23–35.

[50] _____, *Turán-type results for partial orders and intersection graphs of convex sets*, Israel J. Math. **178** (2010), 29–50.

[51] Fox, J. and Sudakov, B., *Density theorems for bipartite graphs and related Ramsey-type results*, Combinatorica **29** (2009), 153–196.

[52] Frankl, P. and Wilson, R. M., *Intersection theorems with geometric consequences*, Combinatorica **1** (4) (1981), 357–368.

[53] Fulek, R. and Pach, J., *A computational approach to Conway's thrackle conjecture*, Comput. Geom. **44** (2011), 345–355.

[54] Fulek, R. and Ruiz-Vargas, A., *Topological graphs: empty triangles and disjoint matchings*, in: Proc. 29th Symp. Comput. Geometry (SoCG 2013), ACM Press, New York, 2013, 259–265.

[55] Fulek, R. and Suk, A., *On disjoint crossing families in geometric graphs*, Electronic Notes in Discrete Mathematics **38** (2011), 367–375.

[56] Füredi, Z., *On the number of edges of quadrilateral-free graphs*, J. Combin. Theory, Ser. B **68** (1996), 1–6.

[57] Gallai, T., *Transitiv orientierbare Graphen*, Acta Math. Acad. Sci. Hungar. **18** (1967), 25–66.

[58] Gilmore, P. C. and Hoffman, A. J., *A characterization of comparability graphs and of interval graphs*, Canad. J. Math. **16** (1964), 539–548.

[59] Golumbic, M., *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[60] Golumbic, M., Rotem, D., and Urrutia, J., *Comparability graphs and intersection graphs*, Discrete Math. **43** (1983), 37–46.

[61] Goodman, J. E. and O'Rourke, J., eds., *Handbook of Discrete and Computational Geometry. 2nd edition,* Chapman & Hall/CRC, Boca Raton, 2004.

[62] Janson, S. and Uzzell, A. J., *On string graph limits and the structure of a typical string graph*, arxiv.org/pdf/1403.2911.pdf

[63] Károlyi, G., Pach, J., and Tóth, G., *Ramsey-type results for geometric graphs. I*, Discrete Comput. Geom. **18** (1997), no. 3, 247–255.

[64] Klazar, M. and Marcus, A., *Extensions of the linear bound in the Füredi-Hajnal conjecture*, Adv. in Appl. Math. **38** (2007), 258–266.

[65] Kleitman, D. J. and Rothschild, B. L., *Asymptotic enumeration of partial orders on a finite set*, Trans. Amer. Math. Soc. **205** (1975), 205–220.

[66] Koebe, P., *Kontaktprobleme der konformen Abbildung*, Berichte über die Verhandlungen der Sachsischen Akademie der Wissenschaften, Leipzig, Mathematische-Physische Klasse **88** (1936), 141–164.

[67] Kolman, P. and Matoušek, J., *Crossing number, pair-crossing number, and expansion*, J. Combin. Theory, Ser. B **92** (2004), no. 1, 99–113.

[68] Kővári, T., Sós, V. T., and Turán, P., *On a problem of K. Zarankiewicz*, Colloquium Math. **3** (1954), 50–57.

[69] Kratochvíl, J., *String graphs. II. Recognizing string graphs is NP-hard*, J. Combin. Theory, Ser. B **52** (1991), 67–78.

[70] Kratochvíl, J. and Matoušek, J., *String graphs requiring exponential representations*, J. Combin. Theory Ser. B **53** (1991), 1–4.

[71] Kupitz, Y. S., *Extremal Problems of Combinatorial Geometry*, Lecture Notes Series **53**, Aarhus University, Denmark, 1979.

[72] Kynčl, J., *Ramsey-type constructions for arrangements of segments*, European J. Combin. **33** (2012), no. 3, 336–339.

[73] ———, *Improved enumeration of simple topological graphs*, Discrete Comput. Geom. **50** (2013), 727–770.

[74] Larman, D., Matoušek, J., Pach, J., and Törőcsik, J., *A Ramsey-type result for convex sets*, Bull. London Math. Soc. **26** (2) (1994), 132–136.

[75] Leighton, T., *Complexity Issues in VLSI. Foundations of Computing Series*, MIT Press, Cambridge, MA, 1983.

[76] Lipton, R. J. and Tarjan, R. E., *A separator theorem for planar graphs*, SIAM J. Appl. Math. **36** (2) (1979), 177–189.

[77] Lovász, L., *Perfect graphs*, in: Selected Topics in Graph Theory, vol. 2, Academic Press, London, 1983, 55–87.

[78] ———, *Geometric Representations of Graphs*, manuscript, http://www.cs.elte.hu/~lovasz-geomrep.pdf.

[79] Lovász, L., Pach, J., and Szegedy, M., *On Conway's thrackle conjecture*, Discrete Comput. Geom. **18** (1997), 369–376.

[80] Lovász, L. and Vesztergombi, K., *Geometric representations of graphs*, in: Paul Erdős and His Mathematics, II. Bolyai Soc. Math. Stud. **11**, János Bolyai Math. Soc., Budapest, 2002, 471–498.

[81] Marcus, A. and Tardos, G., *Excluded permutation matrices and the Stanley-Wilf conjecture*, J. Combin. Theory, Ser. A. **107** (2004), 153–160.

[82] ———, *Intersection reverse sequences and geometric applications*, J. Combin. Theory, Ser. A **113** (2006), no. 4, 675–691.

[83] Matoušek, J., *Lectures on Discrete Geometry*, Springer-Verlag, New York, 2002.

[84] ———, *Near-optimal separators in string graphs*, Combin. Probab. Comput. **23** (2014), no. 1, 135–139.

[85] Miller, G. L., Teng, S.-H., Thurston, W., and Vavasis, S. A., *Separators for sphere-packings and nearest neighbor graphs*, J. ACM **44** (1997), no. 1, 1–29.

[86] Milnor, J., *On the Betti numbers of real varieties*, Proc. of AMS **15** (1964), 275–280.

[87] Pach, J., ed., *Towards a Theory of Geometric Graphs. Contemp. Math.* **342**, Amer. Math. Soc., Providence, RI, 2004.

[88] ———, *Thirty Essays on Geometric Graph Theory*, Springer, New York, 2013.

[89] Pach, J. and Agarwal, P. K., *Combinatorial Geometry*, John Wiley & Sons, New York, 1995.

[90] Pach, J., Pinchasi, R., Sharir, M., and Tóth, G., *Topological graphs with no large grids*, Graphs and Combinatorics **21** (2005), 355–364.

[91] Pach, J., Pinchasi, R., Tardos, G., and Tóth, G., *Geometric graphs with no self-intersecting path of length three*, European J. Combin. **25** (2004), 793–811.

[92] Pach, J., Radoičić, R., and Tóth, G., *Relaxing planarity for topological graphs*, in: Discrete and Computational Geometry, Lecture Notes in Comput. Sci. **2866**, Springer, Berlin, 2003, 221–232.

[93] Pach, J., Shahrokhi, F., and Szegedy, M., *Applications of the crossing number*, Algorithmica **16** (1996), 111–117.

[94] Pach, J. and Solymosi, J., *Crossing patterns of segments*, J. Combin. Theory, Ser. A **96** (2001), 316–325.

[95] Pach, J. and Sterling, E., *Conway's conjecture for monotone thrackles*, Amer. Math. Monthly **118** (2011), 544–548.

[96] Pach, J. and Tóth, G., *Comment on Fox news*, Geombinatorics **15** (2006), no. 3, 150–154.

[97] _____, *How many ways can one draw a graph?*, Combinatorica **26** (2006), no. 5, 559–576.

[98] _____, *Disjoint edges in topological graphs*, J. Comb. **1** (2010), 335–344.

[99] Pach, J. and Törőcsik, J., *Some geometric applications of Dilworth's theorem*, Discrete Comput. Geom. **12** (1994), 1–7.

[100] Pawlik, A., Kozik, J., Krawczyk, T., Lasoń, M., Miczek, P., Trotter, W., and Walczak, B., *Triangle-free intersection graphs of line segments with large chromatic number*, J. Combin. Theory, Ser. B, to appear.

[101] Petrovsky, I. B. and Oleĭnik, O. A., *On the topology of real algebraic surfaces*, Izvestiya Akademii Nauk SSSR**13** (1949), 389–402.

[102] Pinchasi, R. and Radoičić, R., *Topological graphs with no self-intersecting cycle of length* 4, in: Towards a Theory of Geometric Graphs, Contemp. Math. **342**, Amer. Math. Soc., Providence, RI, 2004, 233–243.

[103] Prömel, H. J. and Steger, A., *Excluding induced subgraphs III: A general asymptotic*, Random Structures and Algorithms **3** (1992), 19–31.

[104] Rok, A. and Walczak, B., *Outerstring graphs are χ-bounded*, in: Proc. 30th Symp. Comput. Geometry (SoCG 2014), to appear.

[105] Schaefer, M., *Complexity of some geometric and topological problems*, in: Graph Drawing, Lecture Notes in Comput. Sci. **5849**, Springer, Berlin, 2010, 334–344.

[106] Schaefer, M., Sedgwick, E., and Štefankovič, D., *Recognizing string graphs in NP*, J. Comput. System Sci. **67** (2003), no. 2, 365–380.

[107] Sharir, M. and Agarwal, P. K., *Davenport-Schinzel Sequences and Their Geometric Applications*, Cambridge University Press, Cambridge, 1995.

[108] Sinden, F. W., *Topology of thin film RC-circuits*, Bell System Technical J. **45** (1966), 1639–1662.

[109] Solymosi, J. and Tóth, Cs., *Distinct distances in the plane*, Discrete Comput. Geom. **25** (2001), 629–634.

[110] Suk, A., *Disjoint edges in complete topological graphs*, Discrete Comput. Geom. **49** (2013), no. 2, 280–286.

[111] _____, *Coloring intersection graphs of x-monotone curves in the plane*, Combinatorica, to appear.

[112] _____, *A note on order-type homogeneous point sets*, Mathematika **60** (2014), no. 1, 37–42.

[113] Suk, A. and Walczak, B., *New bounds on the maximum number of edges in k-quasi-planar graphs*, in: Graph Drawing, Lecture Notes in Computer Science **8242**, Springer-Verlag, 2013, 95-106.

[114] Székely, L. A., *Crossing numbers and hard Erdős problems in discrete geometry*, Combin. Probab. Comput. **6** (1997), 353–358.

[115] Tardos, G., *Construction of locally plane graphs with many edges*, in: Thirty Essays on Geometric Graph Theory (J. Pach, ed.), Springer, New York, 2013, 541–562.

[116] Tardos, G. and Tóth, G., *Crossing stars in topological graphs*, SIAM J. Discrete Math. **21** (2007), 737–749.

[117] Thom, R., *Sur l'homologie des variétés algébriques réelles*, in: Differential and Combinatorial Topology (Cairns, S. S., ed.), Princeton University Press, Princeton, NJ, 1965, 255–265.

[118] Tietze, H., *Über das Problem der Nachbargebiete im Raum*, Monatshefte Math. **16** (1905), 211–216.

[119] Tóth, G., *Note on geometric graphs*, J. Combin. Theory, Ser. A **89** (2000), 126–132.

[120] Valtr, P., *Graph drawing with no k pairwise crossing edges*, in: Graph Drawing, Lecture Notes in Comput. Sci. **1353**, Springer, Berlin, 1997, 205–218.

[121] ———, *On geometric graphs with no k pairwise parallel edges*, Discrete Comput. Geom. **19** (1998), 461–469.

[122] Walczak, B., *Triangle-free intersection graphs of line segments with no large independent sets*, manuscript.

[123] Wenger, R., *Helly-type theorems and geometric transversals*, in: Handbook of Discrete and Computational Geometry, CRC Press, Boca Raton, FL, 1997, 63–82.

[124] Woodall, D. R., *Thrackles and deadlock*, in: Combinatorics, Proc. Conf. Comb. Math. (D. Welsh, ed.), Academic Press, London, 1971, 335–347.

École Polytechnique Fédérale de Lausanne, DCG, Station 8, CH-1015 Switzerland;
Rényi Institute, H-1364 Budapest, POB 127, Hungary
E-mail: pachjanos@gmail.com

The determinism of randomness and its use in combinatorics

Angelika Steger

Abstract. Many areas of science, most notably statistical physics, rely on the use of probability theory to explain key phenomena. The aim of this article is to explore the role of probability in combinatorics. More precisely, our aim is to cover a wide range of topics that illustrate the various roles that probability plays within combinatorics: from just providing intuition for deterministic statements, like Szemerédi's regularity lemma or the recent container theorems, over statements about random graphs with structural side constraints and average case analysis of combinatorial algorithms, all the way to neuroscience.

Mathematics Subject Classification (2010). Primary 05C80; Secondary 05A16.

Keywords. Random graph theory, probabilistic methods, extremal graphs, average cases analysis, percolation.

1. Introduction

Probability theory and probabilistic arguments play an important role in many branches of science. The flavor of the probabilistic arguments, however, is often quite different. In statistical physics, for example, a mean field approximation is often the method of choice. Risk and portfolio management aim at modeling the probability distribution of market prices that stem from a badly understood ground truth, while biology relies on the application of statistical methods for designing and interpreting experiments with small sample sizes. In combinatorics, the topic of the present article, the use of probabilistic arguments is again different. Here we are interested in large structures that nevertheless have finite size. And we often use probabilistic arguments (or just intuitions) to derive deterministic or almost deterministic statements. The goal of this article is to provide a variety of such examples that outline a wide range of different uses of probabilistic arguments in combinatorics. We do not aim at being complete, instead we restrict our focus to applications that we will cover in our talk.

We start with two examples. The so-called *probabilistic method* is a nonconstructive method pioneered by Paul Erdős for providing a proof of the existence of certain objects. The methods works by defining an appropriate probability space and then showing that if one chooses a random object from this space, the probability that the result is of the predefined kind is non-zero. As an illustration consider the so-called Ramsey problem: for given k we denote by $R(k)$ the minimal integer n with the property that every 2-coloring of the edges of the complete graph on n vertices contains a monochromatic clique on k vertices. F.P. Ramsey showed that $R(k)$ is finite and Erdős-Szekeres showed inductively that

Proceedings of the International Congress of Mathematicians, Seoul, 2014

$R(k) \leq \binom{2k-2}{k-1}$. To show that $R(k)$ is larger than some integer n we need to construct a 2-coloring of K_n without a monochromatic K_k. As it turns out this is a hard problem - at least if we want to describe (construct) the 2-coloring explicitly. Proving the existence is much easier. Consider a random 2-coloring in which we color each edge red (blue) with probability $1/2$ independently of the other edges. Then the expected number of monochromatic cliques is $\binom{n}{k} \cdot 2 \cdot 2^{-\binom{k}{2}}$. Clearly, if this value is less than 1, there has to *exist* a 2-coloring without monochromatic clique. Straightforward calculations show that this implies $R(k) \geq 2^{(k-1)/2}$. This simple bound is essentially (i.e., up to constant factors) still the best lower bound known.

As a second example consider the so-called Erdős-Rényi random graph model $G_{n,p}$. Here we construct a random graph on n vertices by inserting an edge between any two vertices independently with (edge-)probability p. Clearly, the expected number of edges is $\binom{n}{2}p$ and Chernoff bounds imply that with high probability the actual number of edges is with high probability close to the expected value. The area of random graph theory studies properties of random graph $G_{n,p}$. As it turns out for all reasonable graph properties $\mathcal{P}$ there exists a threshold property $p_0(\mathcal{P})$ such that

$$\text{Prob}[G_{n,p} \text{ has property } \mathcal{P}] = \begin{cases} 1 - o(1) & \text{if } p \ll p_0(\mathcal{P}) \\ o(1) & \text{if } p \gg p_0(\mathcal{P}) \end{cases}$$

(or vice versa). That is, the behavior of a random graph $G_{n,p}$ with respect to property $\mathcal{P}$ is essentially determined by a single parameter, the edge probability p. In this article we will see several examples of such phenomena.

Outline of the paper. In Section 2 we consider classical problems in extremal graph theory and enumerative combinatorics. In particular, we introduce two theorems that contributed dramatically to the progress over the last decades: Szemerédi's regularity lemma at the end of the last century and, more recently, the hypergraph container theorems. Both theorems are deterministic statements. However, they can easily be explained by considering random graphs. Here random graph theory thus provides the intuition for what we aim at. In Section 3 we will use the insights obtained in the first previous section in order to show that and why random instances are often not a good measurement for the algorithmic difficulty of a problem. In Section 4 we move on to extending the classical Erdős-Rényi random graph model to a more sophisticated model of random graphs: that of random *planar* graphs. Over the last decade planar graphs served as one of the role models for studying graph classes where the independence assumption of the edges, as assumed by the Erdős-Rényi random graph model, is not given. Here we will show that the concept of Boltzmann samplers allows to bring in the independence assumption over a 'back door' and we thus can again obtain Chernoff type bounds for many properties. In the last section we will then move on to an application of random graph theory in neuroscience: it is known from empirical observations that locally the brain looks like a random graph.

2. Extremal graph theory and asymptotic counting

One of the very first roots of extremal graph theory, grown long before graph theory per se existed, is the solution to a problem raised by the Dutch mathematician Willem Mantel:

how many edges can a graph have if it does not contain a triangle (that is, a complete graph K_3 on three vertices) as a subgraph? The answer, given by several people, including Mantel himself, shows that the complete bipartite graph with parts as equal as possible is the (unique) extremal graph. The proof of this result is in fact straightforward: consider a triangle-free graph G and let v be a vertex of maximum degree in G. By joining all non-neighbors of v to the neighborhood of v (instead of their current neighbors) we obtain a complete bipartite graph with at least as many edges as G. The result follows. Generalizing Mantel's observation, Paul Turán, a Hungarian mathematician best known for his work in analytic number theory and real and complex analysis, characterized in 1941 the extremal graphs which do not contain the complete graph K_ℓ on ℓ vertices as a subgraph. Turán's result, then, provided stimulation for a variety of problems and results, launching extremal graph theory. His theorem became the most important paradigm for one of the quintessential questions in extremal graph theory: what does an extremal F-free look like? To fix some notation, we denote by $\mathrm{ex}(n, F)$ the maximum number of edges that an F-free graph on n vertices can contain. For $F = K_3$ Mantel's theorem tells us

$$\mathrm{ex}(n, K_3) = \lfloor \frac{n}{2} \rfloor \cdot \lceil \frac{n}{2} \rceil,$$

while for K_ℓ we deduce from Turán's theorem that

$$\mathrm{ex}(n, K_\ell) = \left(1 - \frac{1}{\ell - 1}\right) \frac{n^2}{2} + O(n),$$

where the $O(n)$ term takes care of divisibility issues. In general it turns out that the order of magnitude of $\mathrm{ex}(n, F)$ is always governed by the chromatic number of F. Erdős, Stone [27] and Erdős, Simonovits [28] showed that

$$\mathrm{ex}(n, F) = \left(1 - \frac{1}{\chi(F) - 1} + o(1)\right) \frac{n^2}{2}.$$

There are various ways for proving the Erdős, Stone, Simonovits result. A by now state of the art approach uses Szemerédi's regularity lemma. Here we just give a short sketch of the main idea. A typical property of a random graph is that, with high probability, it has the same density everywhere. The regularity lemma uses this idea and transfers it into a deterministic setting: it is shown that every (dense) graph can be partitioned into a finite number of parts so that almost all of the induced bipartite graphs are 'regular', where the definition of 'regular' captures the features that we know from random graphs: the density should be 'similar' on all sufficiently large subsets. In order to prove the Erdős, Stone, Simonovits result one considers the so-called cluster graph in which every class of the partition is replaced by a vertex and two of these vertices are connected by an edge if the corresponding bipartite graph is regular and sufficiently dense. It is then straightforward to show that the cluster graph has to be $K_{\chi(F)}$-free, which together with Turán's theorem implies the desired bound on the number of edges.

Let $f(n, F)$ denote the number of (labeled) F-free graphs on n vertices. As every subgraph of an F-free graph is also F-free, we trivially have $f(n, F) \geq 2^{\mathrm{ex}(n,F)}$. Erdős, Kleitman and Rothschild [26] showed that in the case of cliques, i.e., for $F = K_\ell$, this lower bound actually provides the correct order of magnitude of the exponent. Erdős, Frankl and Rödl [25] later showed that a similar result holds for all graphs F of chromatic number $\chi(F) \geq 3$:

$$f(n, F) = 2^{(1+o(1))\mathrm{ex}(n,F)}. \tag{2.1}$$

Note that these results provide the correct asymptotics of $\log_2(f(n,F))$. Extending an earlier result from [26] for triangles, Kolaitis, Prömel and Rothschild [38] determined the typical structure of K_ℓ-free graphs by showing that almost all of them are $(\ell-1)$-colorable. Thus,

$$f(n,K_\ell) = (1+o(1)) \cdot \mathrm{col}(n,\ell-1), \tag{2.2}$$

where $\mathrm{col}(n,\ell)$ denotes the number of (labeled) ℓ-colorable graphs on n vertices. An asymptotic for $\mathrm{col}(n,\ell)$ is given in [52]. In the next section we show that this result in fact implies that graph coloring is typically easy in expectation.

Over the last decades the above results were extended in various directions. For example, Prömel and Steger [49] generalized the Kolaitis, Prömel, Rothschild to all graphs with a color-critical edge (an edge is color-critical if its removal reduces the chromatic number of the graph) and Balogh et al [4, 5] improved the error bound in the Erdős, Frankl and Rödl result and provided structural results of a typical F-free graph. In an additional line of research the above results were carried over to the induced case and to hereditary properties in general, see [1, 6, 8, 13, 48, 51] and the references therein. There are, however, two areas where only little progress was made: the case of forbidden bipartite graphs and the case of forbidden graphs of growing size. Here the situation changed only recently with the breakthrough papers of Balogh and Samotij [10, 11] that essentially solved the problem for complete bipartite graphs. In addition, the methods from this paper opened up the way to the development of a general method (nowadays called hypergraph container theorems) that allowed to settle many more important open problems and conjectures.

The plan for the remainder of this section is to outline the recent development on hypergraph container results of Balogh, Morris, Samotij [9] and Saxton and Thomason [53] and to show that they allow to extend the Erdős, Kleitman, Rothschild result to cliques of growing size [44].

We start by outlining the connection of the study of F-free graphs to hypergraph containers. Assume we want to study F-free graphs on vertex set $[n]$. We build an $e(F)$-uniform hypergraph H as follows. The vertex set of H consists of all *edges* of the complete graph on $[n]$, while the edge set consists of all copies of F in the complete graph. That is

$$V(H) = \binom{[n]}{2} \quad \text{and} \quad E(H) = \{X \subset \binom{[n]}{2} \,:\, |X| = e(F) \text{ and } X \cong F\}.$$

Observe that independent sets in H correspond exactly to F-free graphs on vertex set $[n]$ and vice versa. So here is the plan for counting F-free graphs: we apply a container theorem. A container theorem for a hypergraph H is a statement of the following type: there exists a set $\mathcal{C}$ of containers (where a container simply denotes a subset of the vertex set of H) such that three properties are satisfied (for an arbitrary but fixed constant $\varepsilon > 0$):

(i) $\forall I \subset V(H)$ s.t. I is independent: $\exists C \in \mathcal{C}$ s.t. $I \subset C$; i.e., every independent set in H is contained in some container.

(ii) $\forall C \in \mathcal{C} : |E(H[C])| \le \varepsilon |E(H)|$; i.e., the subgraph induced by a container contains at most an ε-fraction of all edges of H.

(iii) $|\mathcal{C}|$ is "small".

Note that the set of all independent sets in H satisfyes the first two constraints, even with $\varepsilon = 0$. Condition (iii) is thus there to enforce non-trivial results. Indeed, in the context of

counting F-free graphs a few moments of thought show that we trivially have

$$f(n,F) \leq \sum_{C \in \mathcal{C}} 2^{|C|}.$$

That is, bounds on $|\mathcal{C}|$ and on $\max_{C\in\mathcal{C}} |C|$ immediately transfer into a bound on the number of F-free graphs. By the definition of the hypergraph H and property (ii) the later condition is equivalent to bounding the number of edges in a graph that contains at most $\varepsilon\binom{n}{v(F)}$ copies of F. It is well known that this number is bounded by $\mathrm{ex}(n,F) + \delta n^2$, for an appropriate constant δ that can be made small by making ε small. (This follows, for example, from the proof of the Erdős, Stone, Simonovits result that we sketched above.) In order to reprove the Erdős, Frankl and Rödl result we thus just need to get a bound for $|\mathcal{C}|$ in the order of $2^{o(n^2)}$ – and the container theorems of Balogh, Morris, Samotij [9] and Saxton and Thomason [53] in fact provide much better bounds.

To get some feeling for which bounds on $|\mathcal{C}|$ are possible we first look at graphs only. A container theorem for *random* graphs $G_{n,p}$ is easily obtained as follows: consider all vertex sets T of size c/p, where $c = c(\varepsilon)$ is an appropriate constant, and set

$$\mathcal{C} = \{T \cup ([n] \setminus \Gamma(T)) \ : \ T \subset [n], |T| = c/p\}, \tag{2.3}$$

where $[n] = \{1,\dots,n\}$ denotes the vertex set of the graph and $\Gamma(T) := \{v \in [n] \setminus T \ : \ \exists w \in T \text{ s.t. } \{v,w\} \in E\}$ denotes the neighborhood of the set T. One easily checks that $\mathcal{C}$ indeed provides a container for every independent set and that $|\mathcal{C}| \leq n^{c/p}$ is "small". It remains to bound the number of edges in a container. Straightforward calculation show that a.a.s. every container $C \in \mathcal{C}$ satisfies $|C| \leq \varepsilon n$ which easily implies (ii).

For general graphs we have to be more careful: we cannot take *all* subsets T of a certain size, as $[n] \setminus \Gamma(T)$ may be too large resp. may contain too many edges. The main idea is to use for an independent set I not an arbitrary subset T, but to argue that there exists a 'clever choice' whose neighborhood $\Gamma(T)$ is large. Indeed, following an approach from [37] we choose the vertices of T iteratively, in each iteration choosing one that adds as many vertices as possible to the set $\Gamma(T)$. Then we know that after we have chosen $t = |T|$ vertices there exists a d such that $\Gamma(T)$ contains at least td vertices and the graph induced by $[n]\setminus\Gamma(T)$ has maximum degree at most d. If we thus assume that the graph G satisfies some local density constraint like 'every not too small set is not too sparse', then this immediately implies a container theorem.

Now consider k-uniform hypergraphs. Again one easily obtains a container theorem for random hypergraphs by taking a similar approach as in (2.3). It is, however, more enlightening to observe what happens if we fix a set $T_k \subseteq [n]$ and consider the induced $(k-1)$-uniform hypergraph given by the edges that contain at least one vertex from T_k. In this hypergraph we then choose a set T_{k-1} and consider the induced $(k-2)$-uniform hypergraph, and so on until we reach a 2-uniform hypergraph, that is, a graph, where we can proceed as indicated above. For random k-uniform hypergraphs $H_{n,p}$ one easily checks that the density of the $(k-i)$-uniform hypergraphs changes in exactly such a way that in all of these hypergraphs we should choose sets T of size $c/p^{1/(k-1)}$. We then obtain a container as the union of the sets T_i plus the complement of the neighborhood of the last set. In order to move from random hypergraphs to arbitrary ones the main idea is similar: we have to choose the sets T_i in some 'clever' way and make some assumptions on the hypergraph under consideration that allow this approach to work. We refer the reader to the original articles by Balogh, Morris,

Samotij [9] and Saxton and Thomason [53] for details and instead just state the consequence that one obtains in the context of clique-free graphs, cf. [44]:

Theorem 2.1. *For every constant $\delta > 0$ the following holds for all large enough $n \in \mathbb{N}$: for every $3 \le \ell \le (\log n)^{1/4}/2$ there exists a collection $\mathcal{G}$ of graphs of order n such that*

(1) *every K_ℓ-free graph of order n is a subgraph of some $G \in \mathcal{G}$,*

(2) *every $G \in \mathcal{G}$ contains at most $\delta\binom{n}{\ell}/e^\ell$ copies of K_ℓ, and*

(3) *the number $|\mathcal{G}|$ of graphs in the collection satisfies $\log|\mathcal{G}| \le \delta n^2/\ell$.*

For graphs that satisfy the second property we can use a theorem of Lovász and Simonovits [40] on the number of cliques in a graph with a given number of edges in order to deduce that the graph cannot contain too many edges. Straightforward calculations then give [44] that for any sequence $(\ell_n)_{n\in\mathbb{N}}$ of positive integers such that $3 \le \ell_n \le (\log n)^{1/4}/2$ we have

$$\log_2 f(n, K_{\ell_n}) = \left(1 - \frac{1}{\ell_n - 1}\right)\binom{n}{2} + o(n^2/\ell_n). \tag{2.4}$$

To the best of our knowledge this is the first non-trivial bound on the number of graphs for a forbidden subgraph of growing order and chromatic number. The paper [44] has already stimulated further research. For example, in [7] Balogh et al determine the typical structure of such graphs. But many problems are still wide open. The upper bound on ℓ_n in (2.4) is just an artifact of the proof. In fact, it is not unconceivable that a similar statement should hold up to the size of a maximal clique in the random graph $G_{n,1/2}$ which is known to be $(2 + o(1))\log_2 n$. It also seems natural that similar results should hold for all forbidden graphs F whose size and chromatic number is bounded by an appropriate function of n.

3. Graph coloring and average case analysis

The graph coloring problem is defined as follows: given a graph G, we search for a proper coloring which uses as few colors as possible. That is, we want to determine $\chi(G)$ and find a coloring of G with exactly $\chi(G)$ colors.

The problem of coloring a graph using a minimum number of colors is one of the central problems of combinatorial optimization. Many application problems, such as register allocation, scheduling problems and pattern matching, can be expressed as special instances of a graph coloring problem. Unfortunately, we know since the seminal work of Karp [36] that graph coloring is a hard problem. More precisely, Karp showed that for any $k \ge 3$ it is $\mathcal{NP}$-complete to decide whether a given graph G is colorable with k colors. More recent results from approximation theory show that it is even quite hard to just approximate the chromatic number. In particular, Feige and Kilian [29] proved that for every $\epsilon > 0$ the existence of a polynomial time approximation algorithm which approximates the chromatic number within a factor of $n^{1-\epsilon}$ implies that $co\mathcal{RP} = \mathcal{NP}$ (which is believed to be false).

Even restricting the input instances to special cases of graphs often leaves the coloring problem surprisingly hard. For example, if we restrict the input to 3-colorable graphs (that is, the input is a graph for which a coloring with 3 colors is known to *exist*, but no such coloring is given), then the best known polynomial-time algorithm for this problem needs roughly $O(n^{0.2111})$ colors [2].

In the light of these difficulties of finding algorithms that are efficient in the worst case it seems natural to hunt for algorithms that are at least fast 'on average'. In his 1984 paper [54] Wilf did just this – in an extremely surprising way. He showed that for every given natural number $k \geq 3$ there exists an *expected* polynomial-time algorithm which decides for every graph $G = (V, E)$ whether $\chi(G) \leq k$ holds and, if so, colors the graph with $\chi(G)$ colors. The surprising fact is that the expected running time of this algorithm is in fact *constant* if one assumes uniform distribution on the set of all graphs. That is, the algorithm decides k-colorability (correctly!) in an expected running time that is not even sufficient to just read a nontrivial part of the graph!

So surprising and unlikely this result may sound at first sight, it is actually an almost trivial consequence of the properties of random graphs. For simplicity we just consider the case $k = 3$. Recall that we get uniform distribution on the set of all graphs by considering the random graph $G_{n,1/2}$. The expected number of K_4's in $G_{n,1/2}$ is $\binom{n}{4}2^{-6} = \Theta(n^4)$. As a copy of a K_4 is a proof that the graph is not 3-colorable, we can thus design an algorithm as follows: consider the vertices $v_1, \ldots, v_n$ sequentially, in every step checking whether v_i forms a K_4 with the vertices $v_1, \ldots, v_{i-1}$. If we find a K_4 we stop and answer 'no', otherwise we check all possible 3-colorings. From (2.1) we know that the probability that the algorithm proceeds to the i-th vertex is bounded by $f(i, K_3) \cdot 2^{-\binom{i}{2}} = 2^{-\Theta(i^2)}$, which immediately implies that the expected running time is constant.

Wilf's result is due to the fact that most graphs contain lots of certificates for the fact that the graph is not 3-colorable. However, we can also turn this around: Prömel and Steger [50] showed that there exists a coloring algorithm $\mathcal{A}$ whose expected running time with respect to the uniform distribution on the class of all $K_{\ell+1}$-free graphs on n vertices is bounded by $O(n^2)$. Note that in this case we know from the Kolaitis, Prömel, Rothschild result [38] that almost all inputs are ℓ-colorable. Here is a sketch of the algorithm. From [38] we know that with high probability a $K_{\ell+1}$-free graphs is typically a random subgraph of the ℓ-partite Turán graph, which implies that for $2 \leq s < \ell$ every K_s-clique can be extended to a K_{s+1}-clique. The coloring algorithm thus works as follows: start with an arbitrary edge and obtain in linear time an ℓ-clique K_ℓ. Now use the 'randomness' property to deduce that there exist many vertices that are connected to $\ell - 1$ vertices of the clique. Observe that if the graph is ℓ-colorable, the coloring of these vertices is determined. In a second round we use the 'randomness' property again to argue that once we have 'many' vertices in each color then this determines with high probability the color of all remaining vertices.

From a mathematical point one may view these kind of results as elegant or significant. Unfortunately, from a practical point of view their value is very limited. Essentially, results of this kind just mean that random instances often do not capture the hardness of a problem appropriately. Within computer science a structural theory of average case complexity was introduced by Levin [39]. We refer the interested reader to the enlightening article of Impagliazzo [31] for a detailed discussion of this topic from a computer science point of view.

We close this section with some remark on the recent progress, cf. [43, 55] and the references therein, on the structure of the solution space of the coloring problem in sparse random graphs. Consider the problem of k-coloring a random graph $G_{n,p}$ where $p = c/n$. If c is small enough there exist 'many' k-colorings, while for c large there will be none with high probability. In order to study this transition one considers the structure of the so-called solution space: the set of all proper k-colorings. We call two solutions close or connected if they can be obtained by changing the coloring of just a constant number of vertices. This

notion of connectedness thus provides a structure on the solution space and we can study how it changes if we increase the average degree of the underlying random graph. As it turns out, first all solutions are in a single cluster. Then clusters of solutions appear but the single giant cluster still exists and dominates. At the so-called clustering transition, the solution space then splits into an exponential number of clusters. A bit later the freezing transition takes place. Here the dominating clusters (that cover almost all proper colorings) start to contain frozen vertices, i.e. vertices that have the same color in all the colorings of the cluster. Understanding (and locating) theses thresholds may well also provide a proof that and why instances 'on the threshold' are computationally hard.

4. Random planar graphs and Boltzmann samplers

The random graph model introduced by Erdős and Rényi gave rise to a beautiful theory with deep theorems and challenging problems whose solution required the development of many new methods and techniques. Unfortunately, intensive studies of real world networks led to the conclusion that the Erdős-Rényi random graph model is often not very well suited. To study real world phenomena one needs more sophisticated models, cf. e.g. the excellent book by Easley and Kleinberg [23] for more information. From a mathematical point of view this calls for the development of new techniques that allow the analysis of random structures that come from a class where the independence assumption between the edges is not given. A typical class with such a property are planar graphs. Clearly, here we have a strong (and non-local) dependence between the presence of edges.

The study of random planar graphs was initiated by Denise, Vasconcellos, and Welsh [18]. In subsequent work McDiarmid, Steger, and Welsh [42] showed that a random planar graph P_n (a graph drawn uniformly at random from the class of all labelled planar graphs on n vertices) in fact has some properties that are very different from the behavior of a classical random graph in the Erdős-Rényi model. In particular, they showed that the probability that P_n is connected is, for n tending to infinity, bounded away from 0 and from 1. This shows that the model of random planar graphs indeed exhibit a behavior that differentiates them strongly from the classical Erdős-Rényi model that has a 0-1-law.

Over the last years two research groups independently developed completely different sets of methods and techniques for attacking this problem. On the one hand a group around Drmota and Noy extended the methods from [30] to develop a framework using techniques from analytic combinatorics. On the other hand Panagiotou and Steger extended the concept of Boltzmann samplers (originally introduced by Duchon et al. [22] for the uniform *generation* of objects), so that it can be used for analysing the *structure* of random planar graphs.

The concept of Boltzmann samplers is algorithmic in nature, but differs from the classical algorithmic approach in that it does not provide an object of a given size but instead provides an object of some size. The Galton-Watson trees are a typical example. Here we start with a root and recursively add descendants according to independent Poisson distributions. To generalize from trees to more general graph classes one can proceed similarly in spirit, but one has to differentiate between the connectivity of the objects under consideration, which adds considerable technical challenges. More precisely, in order to generate connected objects we start with a root and then recursively add for every vertex a certain number of 2-connected objects (blocks) according to independent Poisson distributions. For

the generation of the 2-connected objects we can again use this approach and generate them recursively out of 3-connected objects. The details, however, get much more complicated, as we have to glue along edges instead of vertices.

As it turned out, the block structure of a typical member of a given graph class can be very different [45]. For certain graph classes, like outerplanar and series-parallel graphs, all blocks have only logarithmic size, while for other graphs, like for example planar graphs, there exists one block of linear size, a few blocks of size n^α for some $0 < \alpha < 1$ and the remaining blocks are again of logarithmic size. With this understanding at hand we see that at least for graph classes of the first type we should be able to get very precise statements about their structure: essentially we just have to understand 'small' blocks and then use Chernoff bounds to collect the asymptotic properties of a large collection of (randomly generated) blocks. In this way we get, for example, the degree distribution within these classes [12], cf. also [19].

To extend these results to random planar graphs requires additional efforts, as one also has to understand the structure of the one block of linear size. In [46] a general framework is given for obtaining the degree-sequence for random connected objects from that of a random 2-connected object, and, similarly, for a random 2-connected object from that of a random 3-connected object. Applied to the class of planar graphs, and using the fact that high probability bounds on the degree sequence of a random 3-connected planar graph were known, see [33], this allows to obtain Chernoff like bounds on the degree sequence of a random planar graph. See also [20] for similar results.

In [21] the two groups mentioned above combined their forces (and techniques) to determine the value of the maximum degree in a random planar graph. Extending a previous result of Reed and McDiarmid [41] they showed that the maximum degree of a random planar graph is $c \log n + o(\log n)$ for a constant $c > 0$ that is computed explicitly.

The concept of Boltzmann samplers allows to obtain a variety of Chernoff type results for local properties. For the study of global properties methods from analytic combinatorics, i.e. counting, are still the method of choice. Here we just mention [17, 35] as two examples for some of the recent progress.

5. Percolation theory and neuroscience

How does our brain work? Why can young children perform tasks that renown computer scientists can only dream of to realize? So far, neuroscientists have no answers to these questions. In particular, they do not know how the brain 'computes'. But, of course, they do have a vast knowledge on the physical structure of the brain. One such fact is that, locally, the brain looks like an Erdős-Rényi random graph [34]. Another, known feature of the brain is that information is represented by patterns of activity occurring over populations of neurons [47]. However, so far, there is little understanding how this joint activity of groups of neurons can happen. Experimental observations show that there exists a phenomenon called input normalization (cf. [15] for a review): external input to a local ensemble of neurons (for example from sensory nerves) initially leads only to a small level of activity, which is then boosted by local connectivity. Nevertheless, the total activity never surpasses a certain level. The actual realization of input normalization in the brain is an important topic of research. In this section we argue that percolation theory together with the randomness property of the brain may well provide an answer.

Percolation theory goes back to the pioneering work of Broadbent and Hammersley [14] and has been studied intensively by mathematicians and physicists ever since. A classical and well-studied framework for information spreading is a process known as *bootstrap percolation*. One starts with an initial set of informed or *active* vertices. The process then proceeds in rounds, and further vertices become active as soon as they have at least k active neighbors, where $k \in \mathbb{N}$ is a parameter of the process. This process was first studied in 1979 on the grid by Chalupa and Leath [16] and recently a complete solution for an arbitrary number of dimensions was presented by Balogh, Bollobás, Duminil-Copin and Morris [3]. Janson et al [32] analysed it for the Erdős-Rényi random graph model $G_{n,p}$. A recurring phenomenon in bootstrap percolation theory is a threshold behavior: if the size of the starting set is smaller than some threshold value, we see essentially no percolation at all. If on the other hand the starting set is only slightly above the threshold value we percolate almost completely.

For many percolation processes in physics and material sciences such total percolation is desired and consistent with observations in nature. Not so in neurobiology. Here we have input normalization – and inhibitory neurons that are believed to realize this. In order to obtain a theoretical model for input normalization we extend in [24] the classical bootstrap percolation process in an Erdős-Rényi random graph model $G_{n,p}$ by adding inhibitory vertices. That is, for some constant $\tau \geq 0$ each vertex will be *inhibitory* with probability τ independently of all others. Otherwise the vertex is called excitatory. In each round a previously inactive vertex turns active if the number of active excitatory neighbors exceeds the number of active inhibitory neighbors by at least k.

Unfortunately, it turns out [24] that inhibition has basically no effect on percolation until (possibly) the very end of the process. More precisely, inhibition does not affect the percolation threshold, and it does not prevent (or even slow down) percolation up to an active set of size of $\Omega(1/p)$, a point from which the model without inhibition needs at most two more rounds to activate everything. The reason is that while $o(1/p)$ vertices are active, the typical vertex does not have active inhibitory neighbors, and thus it is not affected by inhibition. On the other hand, if the size of the active set is $\omega(1/p)$, then the behavior of the process depends on the number of inhibitory vertices, i.e. the probability τ. If $\tau < 1/2$ then the process percolates completely, with the same speed as it does without inhibition. Otherwise the process is *chaotic*: tiny variations of the size of the initially active set can have an enormous effect on the size of the final active set. Such a round-based model of percolation can thus not explain the desired phenomenon. In contrast, if we move to a continuous time model in which every edge draws its transmission time randomly, then normalization is an automatic and intrinsic property of the process that holds with very high probability.

References

[1] N. Alon, J. Balogh, B. Bollobás, and R. Morris, *The structure of almost all graphs in a hereditary property*, J. Combin. Theory Ser. B **101** (2011), 85–110.

[2] S. Arora, E. Chlamtac, and M. Charikar, *New approximation guarantee for chromatic number*, Proc. 38th ACM Symposium on Theory of Computing (STOC), 2006, pp. 215–224.

[3] B. Balogh, J.and Bollobás, H. Duminil-Copin, and R. Morris, *The sharp threshold for bootstrap percolation in all dimensions*, Trans. Amer. Math. Soc. **364** (2012), 2667–2701.

[4] J. Balogh, B. Bollobás, and M. Simonovits, *The number of graphs without forbidden subgraphs*, J. Combin. Theory Ser. B **91** (2004), 1–24.

[5] ———, *The typical structure of graphs without given excluded subgraphs*, Random Structures Algorithms **34** (2009), 305–318.

[6] J. Balogh, B. Bollobás, and D. Weinreich, *The speed of hereditary properties of graphs*, J. Combin. Theory Ser. B **79** (2000), 131–156.

[7] J. Balogh, N. Bushaw, M.C. Neto, H. Liu, R. Morris, and M. Sharifzadeh, *The typical structure of graphs with no large cliques*, in preparation.

[8] J. Balogh and J. Butterfield, *Excluding induced subgraphs: critical graphs*, Random Structures Algorithms **38** (2011), 100–120.

[9] J. Balogh, R. Morris, and W. Samotij, *Independent sets in hypergraphs*, J. Amer. Math. Soc., to appear.

[10] J. Balogh and W. Samotij, *The number of $K_{m,m}$-free graphs*, Combinatorica **31** (2011), 131–150 (English).

[11] ———, *The number of $K_{s,t}$-free graphs*, J. Lond. Math. Soc. (2) **83** (2011), 368–388.

[12] N. Bernasconi, K. Panagiotou, and A. Steger, *On properties of random dissections and triangulations*, Combinatorica **30** (2010), 627–654.

[13] B. Bollobás, *Hereditary properties of graphs: asymptotic enumeration, global structure, and colouring*, Proceedings of the International Congress of Mathematicians, Vol. III (Berlin, 1998), Doc. Math. 1998, Extra Vol. III, pp. 333–342 (electronic).

[14] S.R. Broadbent and J.M. Hammersley, *Percolation processes: I. Crystals and Mazes*, Proceedings of the Cambridge Philosophical Society **53** (1957), 629–641.

[15] M. Carandini and D.J. Heeger, *Normalization as a canonical neural computation.*, Nature Reviews Neuroscience **13** (2012), 51–62.

[16] J. Chalupa, P.L. Leath, and G.R. Reich, *Bootstrap percolation on a bethe lattice*, J. Phys. C: Solid State Phys. **12** (1979), L31–L35.

[17] G. Chapuy, E. Fusy, O. Giménez, and M. Noy, *On the diameter of random planar graphs*, preprint, arXiv:1203.3079.

[18] A. Denise, M. Vasconcellos, and D.J.A. Welsh, *The random planar graph*, Congr. Numer. **113** (1996), 61–79.

[19] M. Drmota, O. Giménez, and M. Noy, *Vertices of given degree in series-parallel graphs*, Random Structures Algorithms **36** (2010), 273–314.

[20] ———, *Degree distribution in random planar graphs*, J. Combin. Theory. Ser. A **118** (2011), 2102–2130.

[21] N. Drmota, O. Giménez, M. Noy, K. Panagiotou, and A. Steger, *The maximum degree of random planar graphs*, J. Lond. Math. Soc. (2), to appear.

[22] P. Duchon, P. Flajolet, G. Louchard, and G. Schaeffer, *Boltzmann samplers for the random generation of combinatorial structures*, Combin. Probab. Comput. **13** (2004), 577–625.

[23] D. Easley and J. Kleinberg, *Networks, crowds, and markets*, Cambridge University Press, 2010.

[24] H. Einarsson, J. Lengler, F. Mousset, K. Panagiotou, and A. Steger, *Bootstrap percolation with inhibition*, in preparation.

[25] P. Erdős, P. Frankl, and V. Rödl, *The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent*, Graphs Combin. **2** (1986), 113–121.

[26] P. Erdős, D. Kleitman, and B.L. Rothschild, *Asymptotic enumeration of K_n-free graphs*, Colloquio Internazionale sulle Teorie Combinatorie (Rome, 1973), Tomo II, Atti dei Convegni Lincei, No. 17, Accademia Nazionale dei Lincei, Rome, 1976, pp. 19–27.

[27] P. Erdős and A.H. Stone, *On the structure of linear graphs*, Bull. Amer. Math. Soc. **52** (1946), 1087–1091.

[28] P. Erdős and M. Simonovits, *A limit theorem in graph theory*, Studia Sci. Math. Hungar. **1** (1966), 51–57.

[29] U. Feige and J. Kilian, *Zero knowledge and the chromatic number*, J. Comput. System Sci. **57** (1998), no. 2, 187–199.

[30] O. Giménez and M. Noy, *Asymptotic enumeration and limit laws of planar graphs*, J. Amer. Math. Soc. **22** (2009), 309–329.

[31] R. Impagliazzo, *A personal view of average-case complexity*, Proceedings of the 10th Annual Structure in Complexity Theory Conference (SCT'95) (Washington, DC, USA), SCT '95, IEEE Computer Society, 1995, pp. 134–147.

[32] S. Janson, T. Łuczak, T. Turova, and T. Vallier, *Bootstrap percolation on the random graph $G_{n,p}$*, Ann. Appl. Probab. **22** (2012), 1989–2047.

[33] D. Johannsen and K. Panagiotou, *Vertices of degree k in random maps*, SODA, 2010, pp. 1436–1447.

[34] N. Kalisman, G. Silberberg, and H. Markram, *The neocortical microcircuit as a tabula rasa*, Proc. Natl. Acad.Sci. USA **102** (2005), 880–885.

[35] M. Kang and T. Łuczak, *Two critical periods in the evolution of random planar graphs*, Trans. Amer. Math. Soc. **364** (2012), 4239–4265.

[36] R. Karp, *Reducibility among combinatorial problems*, Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972), Plenum, New York, 1972, pp. 85–103.

[37] D. J. Kleitman and D. Winston, *On the number of graphs without 4-cycles*, Discrete Math. **41** (1982), 167–172.

[38] Ph. Kolaitis, H.J. Prömel, and B.L. Rothschild, *K_{l+1}-free graphs: asymptotic structure and a* 0-1 *law*, Trans. Amer. Math. Soc. **303** (1987), 637–671.

[39] L. Levin, *Average case complete problems*, SIAM J. Comput. **15** (1986), 285–286.

[40] L. Lovász and M. Simonovits, *On the number of complete subgraphs of a graph II*, Studies in Pure Mathematics: To the memory of Paul Turán (Paul Erdős, ed.), Akadémiai Kiadó, Budapest, 1983, pp. 459–495.

[41] C. McDiarmid and B. A. Reed, *On the maximum degree of a random planar graph*, Combin. Probab. Comput. **17** (2008), 591–601.

[42] C. McDiarmid, A. Steger, and D.J.A. Welsh, *Random planar graphs*, J. Combin. Theory Ser. B **93** (2005), 187–205.

[43] M. Molloy, *The freezing threshold for k-colourings of a random graph*, Proc. 44th ACM Symposium on Theory of Computing (STOC), 2012, pp. 921–930.

[44] F. Mousset, R. Nenadov, and A. Steger, *On the number of graphs without large cliques*, SIAM J. on Discrete Math., to appear.

[45] K. Panagiotou and A. Steger, *Maximal biconnected subgraphs of random planar graphs*, ACM Trans. Algorithms **6** (2010), article no. 31.

[46] ———, *On the degree distribution of random planar graphs*, Proc. 22nd ACM-SIAM Symposium on Discrete Algorithms (SODA), 2011, pp. 1198–1210.

[47] A. Pouget, P. Dayan, and R. Zemel, *Information processing with population codes*, Nature Review Neuroscience **1** (2000), 125– 132.

[48] H.J. Prömel and A. Steger, *Almost all Berge graphs are perfect*, Combin. Probab. Comput. **1** (1992), 53–79.

[49] ———, *The asymptotic number of graphs not containing a fixed color-critical subgraph*, Combinatorica **12** (1992), 463–473.

[50] ———, *Coloring clique-free graphs in linear expected time*, Random Structures Algorithms **3** (1992), 375–402.

[51] ———, *Excluding induced subgraphs III: a general asymptotic*, Random Structures Algorithms **3** (1992), 19–31.

[52] ———, *Random ℓ-colorable graphs*, Random Structures Algorithms **6** (1995), 21–37.

[53] D. Saxton and A. Thomason, *Hypergraph containers*, preprint, arXiv:1204.6595.

[54] H. Wilf, *Backtrack: an $O(1)$ expected time algorithm for the graph coloring problem*, Inform. Process. Lett. **18** (1984), 119–121.

[55] L. Zdeborová and F. Krząkała, *Phase transitions in the coloring of random graphs*, Phys. Rev. E **76** (2007), 031131.

Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland
E-mail: steger@inf.ethz.ch

Combinatorial problems in random matrix theory

Van H. Vu

Abstract. In this survey, we discuss several combinatorial problems in Random Matrix theory. We will present the current status of these problems, together with some key ideas and open questions.

Mathematics Subject Classification (2010). Primary 05D40; Secondary 15B52, 60C05.

Keywords. Random matrices, singularity, rank, determinant, least singular value, inverse theorems.

1. Introduction

The theory of random matrices is a very rich topic in mathematics. Beside being interesting in its own right, random matrices play a fundamental role in various areas such as statistics, mathematical physics, combinatorics, theoretical computer science, etc.

In this survey, we focus on problems of a combinatorial nature. These problems are most interesting when the matrix is sampled from a discrete distribution. The most popular models are:

- (Bernoulli) M_n: random matrix of size n whose entries are i.i.d. Bernoulli random variables (taking values ± 1 with probability $1/2$). This is sometimes referred to as the random sign matrix.
- (Symmetric Bernoulli) M_n^{sym}: random symmetric matrix of size n whose (upper triangular) entries are i.i.d. Bernoulli random variables.
- Adjacency matrix of a random graph. This matrix is symmetric and at position ij we write 1 if ij is an edge and zero otherwise.
- Laplacian of a random graph.

Model of random graphs. We consider two models: Erdös-Rényi and random regular graphs. For more information about these models, see [6, 37].

- (Erdös-Rényi) We denote by $G(n,p)$ a random graph on n vertices, generated by drawing an edge between any two vertices with probability p, independently.
- (Random regular graph) A random regular graph on n vertices with degree d is obtained by sampling uniformly over the set of all simple d-regular graphs on the vertex set $\{1,\dots,n\}$. We denote this graph by $G_{n,d}$.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

It is important to notice that the edges of $G_{n,d}$ are not independent. Because of this, this model is usually harder to study, compared to $G(n,p)$.

We denote by $A(n,p)$ ($L(n,p)$) the adjacency (laplacian) matrix of the Erdös-Rényi random graph $G(n,p)$ and by $A_{n,d}$ ($L_{n,d}$) the adjacency (laplacian) matrix of $G_{n,d}$, respectively.

Notation. In the whole paper, we assume that n is large. The asymptotic notation such as o, O, Θ is used under the assumption that $n \to \infty$. We write $A \ll B$ if $A = o(B)$. c denotes a universal constant. All logarithms have natural base, if not specified otherwise.

2. The singular probability

The most famous combinatorial problem concerning random matrices is perhaps the "singularity" problem. Let p_n be the probability that M_n is singular. Trivially,

$$p_n \geq 2^{-n},$$

as the RHS is the probability that the first two rows are equal.

By choosing any two rows (columns) and also replacing equal by equal up to sign, one can have a slightly better lower bound

$$p_n \geq (4 - o(1))\binom{n}{2}2^{-n} = (\frac{1}{2} + o(1))^n. \tag{2.1}$$

It has been conjectured, for quite sometime, that

Conjecture 2.1 (Singularity Conjecture). $p_n = (\frac{1}{2} + o(1))^n$.

Conjecture 2.1 is still open, but one can formulate even more precise conjectures (see [4]), based on the following belief

Phenomenon I. The dominating reason for singularity is the dependency between a few rows/columns.

It is already non-trivial to prove that $p_n = o(1)$. This was first done by Komlós [40] in 1967 and in Section 3, we will give a short proof of this fact. Later, Komlós (see [6]) found a new proof which gave quantitative bound $p_n = O(n^{-1/2})$. In an important paper, Kahn, Komlós and Szemrédi [39] proved the first exponential bound.

Theorem 2.2. $p(n) \leq .999^n$.

Their arguments were simplified by Tao and Vu in 2004 [69], resulting in a slightly better bound $O(.958^n)$. Shortly afterwards, these authors [70] combined the approach from [39] with the idea of inverse theorems (see [74, Chapter 7] or [55] for surveys) to obtained a more significant improvement

Theorem 2.3. $p(n) \leq (3/4 + o(1))^n$.

With an additional twist, Bourgain, Vu and Wood [9] improved the bound further

Theorem 2.4. $p(n) \leq (\frac{1}{\sqrt{2}} + o(1))^n$.

The method from [9, 70] enables one to deduce bounds on p_n directly from simple trigonometrical estimates. For instance, the $3/4$-bound comes from the fact that

$$|\cos x| \leq \frac{3}{4} + \frac{1}{4}\cos 2x,$$

while the $1/\sqrt{2}$ bound come from

$$|\cos x|^2 = \frac{1}{2} + \frac{1}{2}\cos 2x.$$

[9, Theorem 2.2] provides a formal connection between singularity estimates and trigonometric estimates of this type, which, while not yet solving the Singularity Conjecture, does lead to sharp bounds in other situations, such as the singularity of random matrices with $(0, \pm 1)$ entries.

To conclude this section, let us mention a very useful tool, the Littlewood-Offord-Erdös theorem. Let $\mathbf{v} = \{v_1, \dots, v_n\}$ be a set of n non-zero real numbers and $\xi_1, \dots, \xi_n$ be i.i.d random Bernoulli variables. Define $S := \sum_{i=1}^n \xi_i v_i$ and $p_{\mathbf{v}}(a) = \mathbf{P}(S = a)$ and $p_{\mathbf{v}} = \sup_{a \in \mathbf{Z}} p_{\mathbf{v}}(a)$.

The problem of estimating $p_{\mathbf{v}}$ came from a paper of Littlewood and Offord in the 1940s [47], as a key technical ingredient in their study of real roots of random polynomials. Erdös, improving a result of Littlewood and Offord, proved the following theorem, which we will refer to as the Erdös-Littlewood-Offord small ball inequality; see [55] for an explanation of this name.

Theorem 2.5 (Small ball inequality). *Let $v_1, \dots, v_n$ be non-zero numbers and ξ_i be i.i.d Bernoulli random variables. Then*

$$p_{\mathbf{v}} \leq \frac{\binom{n}{\lfloor n/2 \rfloor}}{2^n} = O(n^{-1/2}).$$

Theorem 2.5 is a classical result in combinatorics and has many non-trivial extensions with far reaching consequences (see [7, 36, 55], [74, Chapter 7] and the references therein).

To give the reader a feeling about how small ball estimates can be useful in estimating p_n, let us expose the rows of M_n one by one from top to bottom. Assume that the first $n-1$ rows are independent and form a hyperplane with normal vector $\mathbf{v} = (v_1, \dots, v_n)$. Conditioned on these rows, the probability that M_n is singular is

$$\mathbf{P}(X \cdot \mathbf{v} = 0) = \mathbf{P}(\xi_1 v_1 + \dots + \xi_n v_n = 0),$$

where $X = (\xi_1, \dots, \xi_n)$ is the last row.

In Section 3, the reader will see an application of Theorem 2.5 that leads to Komlós' original result $p_n = o(1)$. In order to obtain the stronger estimates in Theorems 2.3 and 2.4, one needs to ebstablish Inverse (or structural) Littlewood-Offord theorems, based on the following general principle

Phenomenon II. If $\mathbf{P}(X \cdot \mathbf{v} = 0)$ is relatively large, then the coefficients $v_1, \dots, v_n$ posses a strong additive structure.

These theorems are motivated by inverse theorems of Freiman type in Additive Combinatorics, the discussion of which is beyond the scope of this survey. The interested reader is referred to [55] for a detailed discussion.

3. A simple proof of Komlós' Theorem

Let us start with a simple fact. Here and later, Bernoulli vectors mean vectors with coordinates ± 1.

Fact 3.1. *Let H be a subspace of dimension $1 \leq d \leq n$. Then H contains at most 2^d Bernoulli vectors.*

To see this, notice that in a subspace of dimension d, there is a set of d coordinates which determine the others. This fact implies

$$p_n \leq \sum_{i=1}^{n-1} \mathbf{P}(X_{i+1} \in H_i) \leq \sum_{i=1}^{n-1} 2^{i-n} \leq 1 - \frac{2}{2^n},$$

where H_i is the subspace spanned by the first i vectors.

While this bound is quite the opposite of what we want to proof, notice that the loss comes at the end. Thus, to obtain the desired upper bound $o(1)$, it suffices to show that the sum of the last (say) $\log \log n$ terms contribute at most (say) $\frac{1}{\log^{1/3} n}$. To do this, we will exploit the fact that the H_i are spanned by random vectors. The following lemma implies the theorem via the union bound.

Lemma 3.2. *Let H be the subspace spanned by d random vectors, where $d \geq n - \log \log n$. Then with probability at least $1 - \frac{1}{n}$, H contains at most $\frac{2^n}{\log^{1/3} n}$ Bernoulli vectors.*

We say that a set S of d vectors is *k-universal* if for any set of k different indices $1 \leq i_1, \dots, i_k \leq n$ and any set of signs $\epsilon_1, \dots, \epsilon_n$ ($\epsilon_i = \pm 1$), there is a vector v in S such that the sign of the i_jth coordinate of v matches ϵ_j, for all $1 \leq j \leq k$.

Fact 3.3. *If $d \geq n/2$, then with probability at least $1 - \frac{1}{n}$, a set of d random vectors is k-universal, for $k := \log n/10$.*

To prove this, notice that the failure probability is, by the union bound, at most

$$\binom{n}{k}(1 - \frac{1}{2^k})^d \leq n^k (1 - \frac{1}{2^k})^{n/2} \leq n^{-1}.$$

If S is k-universal, then any non-zero vector v in the orthogonal complement of the subspace spanned by S should have more than k non-zero coordinates (otherwise, there would be a vector in S having positive inner product with v). If we fix such v and let X be a random Bernoulli vector, then by Theorem 2.5,

$$\mathbf{P}(X \in Span(S)) \leq \mathbf{P}(X \cdot v = 0) = O(\frac{1}{k^{1/2}}) \leq \frac{1}{\log^{1/3} n},$$

proving Lemma 3.2 and the theorem.

4. The singular probability: symmetric case

As an analogue, it is natural to estimate p_n^{sym}, the probability that the symmetric matrix M_n^{sym} singular.

This problem was mentioned to the author by G. Kalai and N. Linial (personal conversations) around 2004. To our surprise, at that point, even the analogue of Komĺos' 1967 result was not known. According to Kalai and Linial, the following conjecture was circulated by B. Weiss in the 1980s, although it is quite possible that Komlós had thought about it earlier.

Conjecture 4.1. $p_n^{sym} = o(1)$.

The main difficulty concerning M_n^{sym} is that its rows are no longer independent. In particular, the last row is almost determined by the previous ones. Thus, the row exposing procedure considered in the non-symmetric case is no longer useful.

In [18], Costello, Tao and Vu found a way to circumvent the dependency. It turns out that the right way to build the symmetric matrix M_n^{sym} is not row by row (as for M_n), but corner to corner. In step k, one considers the top left sub matrix of size k. The strategy, following an idea of Komlós [40] is to show that with high probability, the co-rank of this matrix, as k increases, behaves like the end point of a biased random walk on non-negative integers which has a strong tendency to go to the left whenever possible. This leads to a confirmation of Weiss' conjecture.

Theorem 4.2. $p_n^{sym} = o(1)$.

The key technical tool in the proof of Theorem 4.2 is the following (quadratic) variant of Theorem 2.5.

Theorem 4.3. *(Quadratic Littlewood-Offord) Let a_{ij} be non-zero real numbers and ξ_i, $1 \leq i, j \leq n$ be i.i.d Bernoulli random variables. Let Q be the quadratic form $Q := \sum_{1 \leq i,j \leq n} a_{ij}\xi_i\xi_j$.. Then for any value a*

$$\mathbf{P}(Q = a) = O(n^{-1/4}).$$

Let us consider the last step in the process when the $(n-1) \times (n-1)$ submatrix M_{n-1}^{sym} has been built. To obtain M_n^{sym}, we add a random row $X = (\xi_1, \dots, \xi_n)$ and its transpose. Conditioning on M_{n-1}^{sym}, the determinant of the resulting $n \times n$ matrix is

$$\sum_{1 \leq i,j \leq n-1} a_{ij}\xi_i\xi_j + \det M_{n-1},$$

where a_{ij} (up to the signs) are the cofactors of M_{n-1}. If M_n^{sym} is singular, then its determinant is 0, which implies

$$Q := \sum_{1 \leq i,j \leq n-1} a_{ij}\xi_i\xi_j = -\det M_{n-1},$$

which gives ground for an application of Theorem 4.3.

Motivated by the non-symmetric case, it is natural to conjecture

Conjecture 4.4. $p_n^{sym} = (1/2 + o(1))^n$.

The concrete bound from [18] is $n^{-1/8}$, which can be easily improved to $n^{-1/4}$. Costello [15] improved the bound to $n^{-1/2+\epsilon}$ and Nguyen [54] pushed it further to $n^{-\omega(1)}$. The best current bound is $\exp(-n^c)$, for some small constant $c > 0$, due to Vershynin [79]. The proofs of the last three results, among others, made sophisticated use of Inverse Littlewood-Offord type results; see [55] for a survey.

5. Ranks and co-ranks

The singular probability is the probability that the random matrix has co-rank at least one. What about larger co-ranks? Let us use $p_{n,k}$ to denote the probability that M_n has co-rank at least k. It is easy to show that

$$p_{n,k} \geq (\frac{1}{2} + o(1))^{kn}. \tag{5.1}$$

It is tempting to conjecture that this bound is sharp for constants k. In [39], Kahn, Komlós and Szemerédi showed

Theorem 5.1. *There is a function $\epsilon(k)$ tending to zero with k such that*

$$p_{n,k} \leq \epsilon^n.$$

In Bourgain et. al. [9], the authors consider a variant of M_n where the first l rows are fixed and the next $n-l$ are random. Let L be the submatrix defined by the first l row and denote the model by $M_n(L)$. It is clear that corank$M_n(L) \geq$ corank$L$. The authors of [9] showed ([9, Theorem 1.4])

Theorem 5.2. *There is a positive constant c such that*

$$\mathbf{P}(\textit{corank}M_n(L) > \textit{corank}L) \leq (1-c)^n.$$

Let us go back to the symmetric model M_n^{sym} and view it from this new angle, exploiting a connection to Erdös-Rényi random graph $G(n, 1/2)$. One can see that

$$M_n^{sym} = 2A(n, 1/2) - J_n,$$

where J_n is the all-one matrix of size n. (Here we allow $G(n, 1/2)$ to have loops, so the diagonal entries of $A(n, 1/2)$ can be one. If we fix all diagonal entries to be zero, the analysis does not change essentially.) Since J_n has rank one, it follows from Theorem 4.2 that with probablity $1 - o(1)$, $A(n, 1/2)$ has corank at most one.

One can reduce the co-rank to zero by a slightly trickier argument. Consider M_{n+1}^{sym} instead of M_n^{sym} and normalize so that its first row and column are all- negative one. Adding this matrix with J_{n+1}, we obtain a matrix of the form

$$\begin{pmatrix} 0 & 0 \\ 0 & M_n^{sym} + J_n \end{pmatrix}$$

Thus we conclude

Corollary 5.3. *With probability $1 - o(1)$, $\textit{corank}A(n, 1/2) = 0$.*

From the random graph point of view, it is natural to ask if this statement holds for a different density p. It is clear that the answer is negative if p is very small. Indeed, if $p < (1-\epsilon)\log n/n$, then $G(n, p)$ has, with high probability, isolated vertices (see [6, 37]) which means that its adjacency matrix has all zero rows and so is singular. Costello and Vu [16] proved that $\log n/n$ is the right threshold.

Theorem 5.4. *For any constant $\epsilon > 0$, with probability $1 - o(1)$,*

$$\textit{corank}A(n, (1+\epsilon)\log n/n) = 0.$$

For $p < \log n/n$, the co-rank of $A(n,p)$ is no longer zero as mentioned above. The behavior of this random variable is not entirely understood. For the case when $p = c\log n/n$ for some constant $0 < c < 1$, Costello et. al. [17] showed that with probability $1 - o(1)$, the co-rank is determined by small subgraphs, which is consistent with **Phenomenon I**. For example,

Theorem 5.5. *For any constant $\epsilon > 0$ and $(1/2+\epsilon)\log n/n < p < (1-\epsilon)\log n/n$, with probability $1 - o(1)$, corank$A(n,(1+p)$ equals the number of isolated vertices in $G(n,p)$.*

For other ranges of p, one needs to take into account the number of cherries (a cherry is a pair of vertices of degree one with a common neighbor) and the numbers of other small subgraphs. The main result of [17] gives a precise formula for the co-rank in terms of these parameters.

When $p = c/n, c > 1$, $G(n,p)$ consists of a giant component and many small components. It makes sense to focus on the giant one which we denote by $Giant(n,p)$. Since $Giant(n,p)$ has cherries , the adjacency matrix of $Giant(n,p)$ is singular (with high probability). However, if we look at the k-core of $Giant(n,p)$, for $k \geq 3$, it seems plausible that this subgraph has full rank.

Conjecture 5.6. *Let k be a fixed integer at least 3. With probability $1 - o(1)$, the adjacency matrix of the k-core of $Giant(n,p)$ is non-singular.*

Bordenave, Lelarge and Salez [8] proved the following related result

Theorem 5.7. *Consider $G(n,c/n)$ for some constant $c > 0$. Then with probability $(1 - o(1))$,*

$$rank(A(n,c/n)) = (2 - q - e^{-cq} - cqe^{-cq} + o(1))n,$$

where $0 < q < 1$ is the smallest solution of $q = \exp(-c\exp -cq)$.

To conclude this section, let us consider the random regular graph $G_{n,d}$. For $d = 2$, $G_{n,d}$ is just the union of disjoint circles. It is not hard to show that with probability $1 - o(1)$, one of these circles has length divisible by 4, and thus its adjacency matrix is non-singular (in fact, the corank will by $\Theta(n)$ as the number of circles of length divisible by 4 is of this order). The following conjecture is open

Conjecture 5.8. *For any $3 \leq d \leq n/2$, with probability $1 - o(1)$ $A_{n,d}$ is non-singular.*

In a recent paper, Cook [14] proved a variant of this conjecture for random d-regular directed graph with large d ($d = \Theta(n)$).

6. Determinant and Permanent

Let us start with a basic question

How big is the determinant of M_n?

This was actually the real motivation of Komlós' original study, as the titles of [40, 41] suggest. However, his results (and other theorems in Section 2) do not give any non-trivial estimate on $|\det M_n|$, except that $|\det M_n| > 0$ with high probability.

As all rows of M_n have length $\sqrt{n}$, Hadamard's inequality implies that $|\det M_n| \leq n^{n/2}$. It has been conjectured that with probability close to 1, $|\det M_n|$ is close to this upper bound.

Conjecture 6.1. *Almost surely* $|\det M_n| = n^{(1/2-o(1))n}$.

This conjecture is supported by a well-known observation of Turán.

Fact 6.2.

$$\mathbf{E}((\det M_n)^2) = n!.$$

To verify this, notice that

$$(\det M_n)^2 = \sum_{\pi,\sigma \in S_n} (-1)^{\text{sign}\pi + \text{sign}\sigma} \prod_{i=1}^{n} \xi_{i\pi(i)}\xi_{i\sigma(i)}.$$

By linearity of expectation and the fact that $\mathbf{E}(\xi_i) = 0$, we have

$$\mathbf{E}(\det M_n)^2 = \sum_{\pi \in S_n} 1 = n!.$$

It follows immediately by Markov's bound that for any function $\omega(n)$ tending to infinity with n,

$$|\det M_n| \leq \omega(n)\sqrt{n!},$$

with probability tending to 1.

A statement of Girko (the main result of [32, 33]) implies that $|\det M_n|$ is typically close to $\sqrt{n!}$. However, his proof appears to contain some gaps (see [56] for details).
In [69], Tao and Vu established the matching lower bound, confirming Conjecture 6.1.

Theorem 6.3. *With probability* $1 - o(1)$,

$$|\det M_n| \geq \sqrt{n!}\exp(-29\sqrt{n\log n}).$$

We sketch the proof very briefly as it contains a useful lemma.

First view $|\det M_n|$ as the volume of the parallelepiped spanned by n random $\{-1, 1\}$ vectors. This volume is the product of the distances from the $(d+1)$st vector to the subspace spanned by the first d vectors, where d runs from 0 to $n-1$. We are able to obtain a very tight control on this distance (as a random variable), thanks to the following lemma, which can be proved using a powerful concentration inequality by Talagrand [69, 82].

Lemma 6.4. *Let* W *be a fixed subspace of dimension* $1 \leq d \leq n-4$ *and* X *a random* ± 1 *vector. For any* $t > 0$

$$\mathbf{P}(|\text{d}ist(X, W) - \sqrt{n-d}| \geq t+1) \leq 4\exp(-t^2/16). \tag{6.1}$$

The lemma, however, is not applicable when d is very close to n. In this case, we need to make use of the fact that W is random, in a fashion similar to the proof in Section 3.

Lemma 6.4 appears handy in many other studies and can be used to derive other concentration inequalities (such as Hanson-Wright type inequalities for concentration of random quadratic forms); see [82] for more details.

Another natural way to estimate $|\det M_n|$ is to view it as the product of the singular values of M_n. By the Marcheko-Pastur law [5], one knows (asymptotically) most singular values. The main obstacle is that the last few can be very small. Thus, the problem basically boils down to bounding the least singular value from below. This problem was first raised by Goldstine and von Neumann in the 1940s [35] and has been investigated in [22, 57, 59, 70, 76] (see also [48, 61] and the references therein for other works concerning rectangular matrices). In particular, Rudelson and Vershynin [59] proved

Theorem 6.5. *There are constants $C, \epsilon > 0$ such that*

$$\mathbf{P}(\sqrt{n}\sigma_{min}(M_n) \leq t) \leq Ct$$

for all $t \leq (1-\epsilon)^n$, where σ_{min} denotes the least singular value.

Theorem 6.5 can be seen as a strengthening of Theorem 2.2; see [55, 58] for more discussion. The bound is sharp, up to the constant C. In [76] the limiting distribution of $\sqrt{n}\sigma_{min}(M_n)$ was determined, yielding the exact value of C in a smaller range of t and settling a conjecture of Edelman and partially a conjecture by Spielman and Teng [63, Conjecutre 2].

Now we turn to the symmetric model M_n^{sym}. Again, by Hadamard's inequality $|\det M_n^{sym}| \leq n^{n/2}$.

Conjecture 6.6. *With probability $1 - o(1)$*

$$|\det M_n^{sym}| = n^{(1/2-o(1))n}.$$

Turán's identity no longer holds because of a correlation caused by symmetry. However, one can still show

$$\mathbf{E}(\det M_n^{sym})^2 = n^{(1+o(1))n}.$$

On the other hand, proving a lower bound for $|\det M_n|$ was more difficult. The problem of bounding the least singular value from below was solved only recently by Nguyen [53] and Vershynin [79], although, unlike the non-symmetric case, we still do not know the limiting distribution of this parameter. The results by Nguyen and Vershynin, combined with the Wigner semi-circle law, confirm Conjecture 6.6

Theorem 6.7. *With probability $1 - o(1)$*

$$|\det M_n^{sym}| = n^{(1/2-o(1))n}.$$

Let us now turn to the related notion of permanent. Recall the formal definition of the determinant of a matrix M (with entries $m_{ij}, 1 \leq i, j \leq n$)

$$\det M := \sum_{\pi \in S_n} (-1)^{\text{sign}\pi} \prod_{i=1}^{n} m_{i\pi(i)}.$$

The permanent of M is defined as

$$\text{Per}M := \sum_{\pi \in S_n} \prod_{i=1}^{n} m_{i\pi(i)}. \tag{6.2}$$

It is easy to see that Turán's identity still holds, namely

$$\mathbf{E}(\operatorname{Per} M_n)^2 = n!.$$

It suggested that $|\operatorname{Per} M_n|$ is typically $n^{(1/2-o(1))n}$. However, this was much harder to prove. The following conjecture, which can be seen as the permanent variant of Komlós classical result $p_n = o(1)$, was open for quite some time

Conjecture 6.8. $\mathbf{P}(\operatorname{Per} M_n = 0) = o(1)$.

The source of difficulties here is that the permanent, unlike the determinant, does not admit any good geometric or linear algebraic interpretation.

In 2007, Tao and Vu found an entirely combinatorial approach to treat the permanent problem [72], relying on the formal definition (6.2) and making heavy use of martingale techniques from probabilistic combinatorics. They proved

Theorem 6.9. *With probability* $1 - o(1)$

$$|\operatorname{Per} M_n| = n^{(1/2-o(1))n}.$$

The still missing (final) piece of the picture is the symmetric counterpart of Theorem 6.9.

Conjecture 6.10. *With probability* $1 - o(1)$

$$|\operatorname{Per} M_n^{sym}| = n^{(1/2-o(1))n}.$$

Motivated by the singularity problem, it is also interesting to find a strong estimate for the probability that the permanent is zero. The current bound is only polynomial in n.

There are further studies concerning the distributions of $\log|\det M_n|$ and $\log|\det M_n^{sym}|$; see [32, 33, 56, 65] and the references therein.

7. Graph expansion and the second eigenvalue

Let G be a connected graph on n points and A its adjacency matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. If G is d-regular then $\lambda_1 = d$ and by Perron-Frobenius theorem no eigenvalue has larger absolute. A parameter of fundamental interest is

$$\lambda(G) := \max_{|\lambda_i|<d} |\lambda_i|.$$

One can derive many interesting properties of the graph from the value this parameter. The general phenomenon here is

Phenomenon III. If $\lambda(G)$ is significantly less than d, then the edges of G distribute like in a random graph with edge density d/n.

This leads to the important notion of pseudo- or quasi-randomness [12] [2]. A representative fact is the following [3]. Let A, B be sets of vertices and $E(A, B)$ the number of edges with one end point in A and the other in B, then

$$|E(A,B) - \frac{d}{n}|A||B|| \leq \lambda(G)\sqrt{|A||B|}. \tag{7.1}$$

Notice that the term $\frac{d}{n}|A||B|$ is the expectation of the number of edges between A and B if G is random (in the Erdös-Rényi sense) with edge density d/n. Graphs with small λ are often called *pseudo-random* [12, 42].

One can use this information about edge distribution to derive various properties of the graph (see [42] for many results of this kind). The whole concept can be generalized for non-regular graphs, using the Laplacian rather than the adjacency matrix (see, for example, [13]).

From (7.1), it is clear that the smaller λ is, the more "random" G is. *But how small can λ be ?*

Alon and Boppana [1] proved that if d is fixed and n tends to infinity, then

$$\lambda(G) \geq 2\sqrt{d-1} - o(1).$$

Graphs which satisfy $\lambda(G) < 2\sqrt{d-1}$ are called Ramanujan graphs. It is difficult to construct such graphs, and all known constructions, such as those by Lubotzky-Phillip-Sarnak [43] and Margulis [44], rely heavily on number theoretic results, which apply only to specific values of d. A more combinatorial approach was found recently by Markus, Spielman, and Snivastava [49]. Their method (at least in the bipartite case) works for all d, but the construction is not explicit.

Theorem 7.1. *A bipartite Ramanujan graph exists for all fixed degrees $d \geq 3$ and sufficiently large n.*

While showing the existence of Ramanujan graphs is already highly non-trivial, the real question, in our opinion, is to compute the limiting distribution of $\lambda(G_{n,d}) - 2\sqrt{d-1}$ after a proper normalization, which would lead to the exact probability of a random regular graph being Ramanujan. Motivated by studies from Random matrix theory, it seems plausible to conjecture that $n^{2/3}(\frac{\lambda(G(n,d))}{\sqrt{d-1}} - 2)$ tends to the Tracy-Widom distribution.

A weaker conjecture, by Alon [1] asserts that for any fixed d, with probability $1 - o(1)$

$$\lambda_2(G_{n,d}) = 2\sqrt{d-1} + o(1).$$

Friedman [28] and Kahn and Szemerédi [38] showed that if d is fixed and n tends to infinity, then with probability $1 - o(1)$, $\lambda(G_{n,d}) = O(\sqrt{d})$. About 10 years ago, Friedman, in a highly technical paper [29], used the moment method to prove Alon's conjecture (see also [30] for a recent genearlization)

Theorem 7.2 ([29]). *For any fixed d and n tends to infinity, with probability $1 - o(1)$*

$$\lambda(G_{n,d}) = 2\sqrt{d-1} + o(1).$$

What happens if d tends to infinity with n? To start, it is not hard to show that $\lambda(G(n,p))$, where $G(n,p)$ is the Erdös-Réyi random graph, is $(2 + o(1))\sqrt{np(1-p)}$ for sufficiently large p (e.g., $p \geq n^{-1+\epsilon}$ for any fixed $0 < \epsilon < 1$). This motivates

Conjecture 7.3. *Assume that $d \leq n/2$ and both d and n tend to infinity. Then with probability $1 - o(1)$,*

$$\lambda(G_{n,d}) = (2 + o(1))\sqrt{d(1-d/n)}.$$

Nilli [51] showed that for any d-regular graph G having two edges with distance at least $2k+2$ between them, $\lambda_2(G) \geq 2\sqrt{d-1} - 2\sqrt{d-1}/(k+1)$. Any d regular graph with $d = n^{o(1)}$ has diameter $\omega(1)$. In this range of d

$$\lambda(G_{n,d}) \geq \lambda_2(G_{n,d}) \geq (2+o(1))\sqrt{d}$$

with probability one. This proves the lower bound in Conjecture 7.3. For a general d, it is easy to show (by computing the trace of the square of the adjacency matrix) that any d-regular graph G on n vertices satisfies

$$\lambda(G) \geq \sqrt{d(n-d)/(n-1)} \approx \sqrt{d(1-d/n)}.$$

(We would like to thank N. Alon for pointing out this bound.)

Let us now turn to the upper bound. For $d = o(n^{1/2})$, one can follow the Kahn-Szemerédi approach to show that $\lambda(G_{n,d}) = O(\sqrt{d})$ with high probability. However, we do not know this for larger d. For instance, the following is open

Conjecture 7.4. *With probability* $1-o(1)$, $\lambda(G_{n,n/2}) = O(\sqrt{n})$.

8. Eigenvectors

If M is symmetric, then its (unit) eigenvectors form an orthonormal basis. Works concerning random eigenvectors are generally motivated by

Phenomenon IV. *Random eigenvectors should behave like a random vector sampled uniformly from the unit sphere .*

One parameter which has been looked at a lot is the infinite norm, as it plays a big role in recent studies on universality (see [27, 75] for surveys). Following earlier results [26, 67], recently Vu and Wang [82] proved

Theorem 8.1. *With probability* $1-o(1)$,

$$\max \|v\|_\infty \leq C\sqrt{\frac{\log n}{n}},$$

where the maximum is taken over the"bulk" eigenvectors of M_n^{sym}. *If one also considers the "edge" eigenvectors, the bound becomes* $C\frac{\log n}{\sqrt{n}}$, *where* C *is a constant.*

Notice that a vector sampled uniformly from the unit sphere does have a coordinate of magnitude $\Theta(\sqrt{\frac{\log n}{n}})$, we believe that the bound $O(\sqrt{\frac{\log n}{n}})$ is best possible. Similar, but weaker, results (with higher powers of $\log n$) hold for the non-symmetric model M_n, with respect to both singular vectors and eigenvectors [60, 68].

The situation with the adjacency matrix of a random graph is somewhat more complicated. Consider $A(n,p)$ with $p = \Theta(1)$. The sum of any rows is close to np. It suggests that the largest eigenvalue λ_1 of $A(n,p)$ is approximately np and its corresponding eigenvector v_1 is close to $\frac{1}{\sqrt{n}}v_0$, where v_0 is the all-one vector. This intuition was confirmed by Komlós and Füredi [31], and strengthened by Mitra [46].

In [19], Dekel, Lee and Linial, motivated by the study of nodal domains, raised the following question.

Question 8.2. *Is it true that every eigenvector u of $G(n,p)$ has $||u||_\infty = n^{-1/2+o(1)}$ with high probability ?*

For many related results, we refer to [24, 25, 78]. Another question motivated by Phenomenon IV is the following.

Conjecture 8.3. *Assume $p \geq \frac{(1+\epsilon)\log n}{n}$ for some constant $\epsilon > 0$. Let v be a random unit vector whose distribution is uniform in the n-dimensional unit sphere. Let u be a unit eigenvector (not corresponding to the largest eigenvalue) of $G(n,p)$. Then for any fixed $\delta > 0$ and unit vector w*

$$\mathbf{P}(|w \cdot u - w \cdot v| > \delta) = o(1).$$

For related results, see [10, 77].

Let us now consider random regular graphs. Recently Dimitriu and Pal [20] proved the following result. Let $d = \log^\gamma n$ for a constant $0 < \gamma < 1$, and set $\eta_n := \frac{6(\log d)^{1+\sigma}}{\sqrt{\log n}}$ where $\sigma > 0$ is a constant. A unit vector $v = (v_1, \ldots, v_n)$ is (T, ϵ)-localized if there is a set X of size T such that $\sum_{i \in X} v_i^2 \geq \epsilon$.

Theorem 8.4. *For any fixed $\epsilon > 0$, with probability $1 - o(1)$, no eigenvector of $A(n,d)$ is $(o(\eta_n^{-1}), \epsilon)$-localized.*

A more recent result of Brooks and Lindenstrauss [11] showed

Theorem 8.5. *Let d, ϵ be constants. Then there is a constant $\delta = \delta(d, \varepsilon) > 0$ such that the following holds. With probability $1 - o(1)$, no eigenvector of $A(n,d)$ is (n^δ, ϵ) localized.*

In fact, Brooks and Lindenstrauss' result holds for deterministic graphs, under a condition on short cycles, which holds with high probability for regular random graphs with constant degree.

Problem 8.6. Can we replace the (n^δ, ϵ) -localization in Theorem 8.5 by $(\delta n, \epsilon)$-localization?

9. Random regular graphs: Mc Kay law and Wigner law

We briefly discuss the spectral distribution of regular random graphs. In 1950s, Wigner [80] discovered the famous semi-circle law for the limiting distribution of the eigenvalues of random matrices. His proof extends, without difficulty, to the adjacency matrix of $G(n,p)$, given that $np \to \infty$ with n.

Theorem 9.1. *For $p = \omega(\frac{1}{n})$, the empirical spectral distribution (ESD) of the matrix $\frac{1}{\sqrt{n}\sigma} A_n$ converges in distribution to the semicircle law which has a density $\rho_{sc}(x)$ with support on $[-2, 2]$,*

$$\rho_{sc}(x) := \frac{1}{2\pi}\sqrt{4 - x^2}.$$

If $np = O(1)$, the semicircle law no longer holds. In this case, the graph almost surely has $\Theta(n)$ isolated vertices, so in the limit, the origin has a positive constant mass.

The case of random regular graph, $G_{n,d}$, was considered by McKay [45] about 30 years ago. He proved, using the trace method, that if d is fixed, and $n \to \infty$, then the limiting density function is

$$f_d(x) = \begin{cases} \frac{d\sqrt{4(d-1)-x^2}}{2\pi(d^2-x^2)}, & \text{if } |x| \le 2\sqrt{d-1}; \\ \\ 0 & \text{otherwise.} \end{cases}$$

This is usually referred to as McKay or Kesten-McKay law. It is easy to verify that as $d \to \infty$, if we normalize the variable x by $\sqrt{d-1}$, the above density converges to the semicircle law on $[-2, 2]$. It is thus natural to conjecture that Theorem 9.1 holds for $G_{n,d}$ with $d \to \infty$. Define

$$M'_{n,d} := \frac{1}{\sqrt{d}}(A_{n,d} - \frac{d}{n}J).$$

Conjecture 9.2. *If $d \to \infty$ then the ESD of $\frac{1}{\sqrt{n}}M'_{n,d}$ converges to the semicircle law.*

Dimitriu and Pal [20] showed that the conjecture holds for d tending to infinity very slowly, $d = n^{o(1)}$. Their proof which used the trace method does not work for larger d as it relies on the tree-like local structure of the graph, which no longer holds if $d = n^c$ for any constant $c > 0$. Very recently, Tran, Vu and Wang [78] proved Conjecture 9.2 in full generality, using a completely different method using a sharp concentration result from [34].

Theorem 9.3. *If d tends to infinity as n goes to infinity, then the empirical spectral distribution of $\frac{1}{\sqrt{n}}M'_n$ converges in distribution to the semicircle distribution.*

10. Miscellany

About 15 years ago, Krivelevich asked the following question: Is it true that (with probability $1 - o(1)$), $A(n, 1/2)$ does not have any multiple eigenvalues ?

In a more recent conversation, L. Babai mentioned that he came up with the same question much earlier. We strongly believe that the answer to this question is affirmative, and the same must hold for other models of random matrices.

Conjecture 10.1. *With probability $1 - o(1)$,*

- *$A(n, 1/2)$ does not have multiple eigenvalues.*
- *M_n does not have multiple eigenvalues.*
- *M_n does not have multiple singular values.*
- *M_n^{sym} does not have multiple eigenvalues.*
- *M_n^{sym} does not have multiple singular values.*

Another interesting (and seemingly very hard) conjecture is the following, which came up in the conversation between the author and P. Wood in 2009. Recently, L. Babai informed us that he made the same conjecture (unpublished) in the 1970s.

Conjecture 10.2. *With probability $1 - o(1)$, the characteristic polynomial of M_n is irreducible.*

Here is another conjecture

Conjecture 10.3. *A ± 1 matrix is determined by its spectrum if no other ± 1 matrix has the same spectrum. Prove that almost all ± 1 matrices are determined by their spectrum (not counting trivial permutations).*

The following conjecture is motivated by our joint work with Tao in [73]

Conjecture 10.4. *M_n has, with high probability, $\Theta(\sqrt{n})$ real eigenvalues.*

Edelman, Kostlan and Shub [21] obtained a formula for the expectation of the number of real eigenvalues for a gaussian matrix (which is or order $\Theta(\sqrt{n})$). In [73], Tao and Vu proved that the same formula holds (in the asymptotic sense) for certain random matrices with entries $(0, \pm 1)$. However, we do not know anything for M_n. As a matter of fact, even the following "first step" looks non-trivial.

Problem 10.5. Prove that M_n has, with high probability, at least 2 real eigenvalues.

The next problem bears some resemblance to the famous "rigidity" problem in computer science. Given $\{-1, 1\}$ matrix M, we denote by $Res(M)$ the minimum number of entries we need to switch (from 1 to -1 and vice versa) in order to make M singular. If M is a sample of M_n, it is easy to show that $Res(M)$ is, with high probability, at most $(1/2 + o(1))n$, We conjecture that this is the best one can do.

Conjecture 10.6. *With probability $1 - o(1)$, $Res(M_n) = (1/2 + o(1))n$.*

A closely related question (motivated by the notion of local resilience from [64]) is the following. Call a $\{-1, 1\}$ (n by n) matrix M *good* if all matrices obtained by switching (from 1 to -1 and vice versa) the diagonal entries of M are non-singular (there are 2^n such matrices).

Conjecture 10.7. *With probability $1 - o(1)$, M_n is good.*

Finally, let us list a few recent papers concerning groups defined over random matrices with entries from a finite field [50, 83]. This direction is new and these works need an elaborate introduction, which will appear elsewhere.

Acknowledgements. The author would like to thank NSF and AFORS for their generous support and K. Luh for his careful proofreading.

References

[1] N. Alon, *Eigenvalues and expanders*, Combinatorica **6** (1986), no. 2, 83–96.

[2] N. Alon and V. Milman, λ_1- *isoperimetric inequalities for graphs, and supercon- centrators*, J. Combin. Theory Ser. B **38** (1985), no. 1, 73–88.

[3] N. Alon and J. Spencer, *The probabilistic method*, 3rd ed., John Wiley & Sons Inc., Hoboken, NJ, 2008.

[4] R. Arratia and S. DeSalvo, *On the singularity of random Bernoulli matrices-novel integer partitions and lower bound expansions*, Ann. Comb. **17** (2013), no. 2, 251–274.

[5] Z. Bai and J. Silverstein, *Spectral analysis of large dimensional random matrices*, Second edition, Springer Series in Statistics. Springer, New York, 2010.

[6] B. Bollobás, *Random graphs*, Second edition, Cambridge Studies in Advanced Mathematics, 73. Cambridge University Press, Cambridge, 2001.

[7] _____, *Combinatorics*, Set systems, hypergraphs, families of vectors and combinatorial probability, Cambridge University Press, Cambridge, 1986.

[8] C. Bordenave, M. Lelarge, and J. Salez, *The rank of diluted random graphs*, Ann. Probab. **39** (2011), no. 3, 1097–1121.

[9] J. Bourgain, V. Vu, and P. M. Wood, *On the singularity probability of discrete random matrices*, J. Funct. Anal. **258** (2010), no. 2, 559–603.

[10] P. Bourgade and H-T Yau, *The Eigenvector Moment Flow and local Quantum Unique Ergodicty*, arXiv:1312.1301.

[11] S. Brooks and E. Lindenstrauss, *Non-localization of eigenfunctions on large regular graphs*, Israel J. Math. **193** (2013), no. 1, 1–14

[12] Chung, F. R. K., Graham, R. L., and Wilson, R. M., *Quasi-random graphs*, Combinatorica **9** (1989), no. 4, 345–362.

[13] F. Chung, *Spectral graph theory*, CBMS series, no. 92, 1997.

[14] N. Cook, *Random regular graphs: Singularity and Discrepancy*, arXiv:1403.5845.

[15] K. Costello, *Bilinear and quadratic variants on the Littlewood-Offord problem*, Israel J. Math. **194** (2013), no. 1, 359–394.

[16] K. Costello and V. Vu, *The ranks of random graphs*, Random Structures and Algorithm. **33** (2008), 269–285

[17] _____, *The rank of sparse random matrices*, Combin. Probab. Comput. **19** (2010), no. 3, 321–342.

[18] K. Costello, T. Tao, and V. Vu, *Random symmetric matrices are almost surely singular*, Duke Math. J. **135** (2006), no. 2, 395–413.

[19] Y. Dekel, J. Lee, and N. Linial, *Eigenvectors of random graphs: Nodal domains*, Approx- imation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 2008, pp. 436–448.

[20] I. Dimitriu and S. Pal, *Sparse regular random graphs: spectral density and eigenvectors*, Ann. Probab. **40** (2012), no. 5, 2197–2235.

[21] A. Edelman, E. Kostlan, and M. Shub, *How many eigenvalues of a random matrix are real?*, J. Amer. Math. Soc. **7** (1994), no. 1, 247–267.

[22] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl. **9** (1988), 543–560.

[23] P. Erdös, *On a lemma of Littlewood and Offord*, Bull. Amer. Math. Soc. **51** (1945), 898–902.

[24] L. Erdös, A. Knowles, H-T. Yau, and J. Yin, *Spectral statistics of Erdős-Rényi graphs I: Local semicircle law*, Ann. Probab. **41** (2013).

[25] L. Erdös, A. Knowles, H-T. Yau, and J. Yin, *Spectral statistics of Erdős-Rényi graphs II: Eigenvalue spacing and the extreme eigenvalues*, Comm. Math. Phys. **314** (2012), no. 3, 587–640.

[26] L. Erdös, B. Schlein, and H-T. Yau, *Wegner estimate and level repulsion for Wigner random matrices*, Int. Math. Res. Not. IMRN 2010, no. 3, 436–479.

[27] L. Erdös and H-T. Yau, *Universality of local spectral statistics of random matrices*, Bull. Amer. Math. Soc. (N.S.) **49** (2012), no. 3, 377–414.

[28] J. Friedman. *On the second eigenvalue and random walks in random d-regular graphs*, Technical Report CX-TR-172-88, Princeton University, August 1988.

[29] _____, *A proof of Alon's second eigenvalue conjecture and related problems*, (English summary)*Mem. Amer. Math. Soc.* **195** (2008), no. 910, viii+100 pp.

[30] J. Friedman and D-E. Kohler, *The Relativized Second Eigenvalue Conjecture of Alon*, preprint.

[31] Z. Füredi and J. Komlós, *The eigenvalues of random symmetric matrices*,Combinatorica **1** (1981), no. 3, 233–241.

[32] V. L. Girko, *A refinement of the central limit theorem for random determinants*, (Russian) Teor. Veroyatnost. i Primenen. **42** (1997), no. 1, 63–73; translation in Theory Probab. Appl. **42** (1997), no. 1, 121–129 (1998)

[33] _____, *A central limit theorem for random determinants*, Teor. Veroyatnost. i Mat. Statist. **21** (1979), 35–39, 164.

[34] A. Guionnet and O. Zeitouni, *Concentration of the spectral measure for large matrices*, Electron. Comm. Probab. **5** (2000), 119–136.

[35] H. Golstein and J. von Neumman, *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc. **53** (1947), 1021–1099.

[36] G. Halász, *Estimates for the concentration function of combinatorial number theory and probability*, Period. Math. Hungar. **8** (1977), no. 3-4, 197–211.

[37] S. Janson, T. Luczak, and A. Rucinski, *Random Graphs*, Wiley-Interscience (2000)

[38] J. Kahn and E. Szemerédi, STOC 1989.

[39] J. Kahn, J. Komlós, E. Szemerédi, *On the probability that a random* ± 1 *matrix is singular*, J. Amer. Math. Soc. **8** (1995), 223–240.

[40] J. Komlós, *On the determinant of* $(0, 1)$ *matrices*, Studia Sci. Math. Hungar. **2** (1967) 7–22.

[41] _____, *On the determinant of random matrices*, Studia Sci. Math. Hungar. **3** (1968) 387–399.

[42] M. Krivelevich and B. Sudakov, *Pseudo-random graphs*, More sets, graphs and numbers, 199–262, Bolyai Soc. Math. Stud., 15, Springer, Berlin, 2006.

[43] A. Lubotzky, R. Phillips, and P. Sarnak, *Ramanujan graphs*, Combinatorica, **8**(3) (1998), 261–277.

[44] G.A. Margulis, *Explicit group-theoretical constructions of combinatorial schemes and their application to the design of expanders and superconcentrators [in Russian]*, Problemy Peredachi Informatsii **24** (1988), 51–60.

[45] B.D. McKay, *The expected eigenvalue distribution of a large regular graph*, Linear Algebra and its Applications, **40** (1981), 203–216.

[46] P. Mitra, *Entrywise bounds for eigenvectors of random graphs*, Electron. J. Combin. **16** (2009), no. 1, Research Paper 131,

[47] J. E. Littlewood and A. C. Offord, *On the number of real roots of a random algebraic equation. III*, Rec. Math. [Mat. Sbornik] N.S. **12** (1943), 277–286.

[48] A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Adv. Math. **195** (2005), no. 2, 491–523.

[49] A. Marcus, D. Spielman, and N. Srivastava, *Interlacing Families I: Bipartite Ramanujan Graphs of All Degrees*, preprint.

[50] K. Maples, *Symmetric random matrices over finite fields announcement*, April 15, 2013, preprint.

[51] A. Nilli, *On the second eigenvalue of a graph*, Discrete Mathematics **91** (1991), 207–210.

[52] _____, *Tight estimates for eigenvalues of regular graphs*, Electronic J. Combinatorics **11** (2004), N9, 4pp.

[53] H. Nguyen, *On the least singular value of random symmetric matrices*, Electron. J. Probab. **17** (2012), no. 53.

[54] _____, *Inverse Littlewood-Offord problems and the singularity of random symmetric matrices*, Duke Math. J. **161** (2012), no. 4, 545–586.

[55] H. Nguyen and V. Vu, *Small probability, inverse theorems, and applications*, Erdos' 100th Anniversary Proceeding, Bolyai Society Mathematical Studies, Vol. 25 (2013).

[56] _____, *Random matrices: Law of the determinant*, Annals of Probability (2014), Vol. 42, No. 1, 146–167.

[57] M. Rudelson, *Invertibility of random matrices: norm of the inverse*, Ann. of Math. (2) **168** (2008), no. 2, 575–600.

[58] _____, *Lecture notes on non-aymptotic random matrix theory*, notes from the AMS Short Course on Random Matrices, 2013.

[59] M. Rudelson and R. Vershynin, *The Littlewood-Offord problem and invertibility of random matrices*, Adv. Math. **218** (2008), no. 2, 600–633.

[60] M. Rudelson and R. Vershynin, *Delocalization of eigenvectors of random matrices with independent entries*, preprint.

[61] O. N. Feldheim and S. Sodin, *A universality result for the smallest eigenvalues of certain sample covariance matrices*, Geom. Funct. Anal. **20** (2010), no. 1, 88–123.

[62] A. Sárközy and E. Szemerédi, *Uber ein Problem von Erdös und Moser*, Acta Arithmetica, **11** (1965) 205–208.

[63] D. Spielman and S-H. Teng, D. Spielman, and S.-H. Teng, *Smoothed analysis of algorithms*, Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002), 597–606, Higher Ed. Press, Beijing, 2002.

[64] B. Sudakov and V. Vu, *Local resilience of graphs*, Random Structures Algorithms **33** (2008), no. 4, 409–433.

[65] T. Tao and V. Vu, *A central limit theorem for the determinant of a Wigner matrix*, Adv. Math. **231** (2012), no. 1, 74–101.

[66] _____, *Random matrices: universal properties of eigenvectors*, Random Matrices Theory Appl. **1** (2012), no. 1.

[67] _____, *Random matrices: Universality of the local eigenvalues statistics pdf file*, Acta Math. **206** (2011), no. 1, 127–204.

[68] _____, *Random covariance matrices: universality of local statistics of eigenvalues*, Ann. Probab. **40** (2012), no. 3, 1285–1315.

[69] _____, *On random* ± 1 *matrices: Singularity Determinant*, Random Structures Algorithms **28** (2006), no. 1, 1–23.

[70] _____, *On the singularity probability of random Bernoulli matrices*, J. Amer. Math. Soc. **20** (2007), no. 3, 603–628.

[71] _____, *Inverse Littlewood-Offord theorems and the condition number of random matrices*, Annals of Math. **169** (2009), 595-632

[72] _____, *On the permanent of random Bernoulli matrices*,Advances in Mathematics **220** (2009), 657-669.

[73] _____, *Random matrices: Universality of local spectral statistics of non-Hermitian matrices*, to appear in Annals of Probability.

[74] _____, *Additive Combinatorics*, Cambridge Univ. Press, 2006.

[75] _____, *Random matrices: The Universality phenomenon for Wigner ensembles*, preprint, to appear in AMS lecture notes on Random Matrices, 2013.

[76] _____, *Random matrices: the distribution of the smallest singular values*, Geom. Funct. Anal. **20** (2010), no. 1, 260-297.

[77] _____, *Random matrices: universal properties of eigenvectors*, Random Matrices Theory Appl. **1** (2012), no. 1.

[78] L. Tran, V. Vu and K. Wang, *Sparse random graphs: Eigenvalues and Eigenvectors*, Random Structures Algorithms **42** (2013), no. 1, 110–134.

[79] R. Vershynin, *Invertibility of symmetric random matrices*, Random Structures and Algorithms **44** (2014), 135–182

[80] E.P. Wigner, *On the distribution of the roots of certain symmetric matrices*, Annals of Mathematics, **67**(2):325-327, 1958.

[81] N.C. Wormald, *Models of random regular graphs*, In Surveys in Combinatorics, 1999, J.D. Lamb and D.A. Preece, eds, pp. 239-298.

[82] V. Vu and K. Wang, *Random projection, random quadratic forms, and random eigenvectors*, to appear in Random Structures and Algorithms.

[83] M. Wood, *The distribution of sandpile groups of random graphs*, preprint.

Department of Mathematics, Yale University, New Haven, CT, USA
E-mail: van.vu@yale.edu

14. Mathematical Aspects of Computer Science

Sum-of-squares proofs and the quest toward optimal algorithms

Boaz Barak and David Steurer

Abstract. In order to obtain the best-known guarantees, algorithms are traditionally tailored to the particular problem we want to solve. Two recent developments, the *Unique Games Conjecture (UGC)* and the *Sum-of-Squares (SOS) method*, surprisingly suggest that this tailoring is not necessary and that a single efficient algorithm could achieve best possible guarantees for a wide range of different problems. The *Unique Games Conjecture (UGC)* is a tantalizing conjecture in computational complexity, which, if true, will shed light on the complexity of a great many problems. In particular this conjecture predicts that a *single concrete algorithm* provides optimal guarantees among all efficient algorithms for a large class of computational problems. The *Sum-of-Squares (SOS) method* is a general approach for solving systems of polynomial constraints. This approach is studied in several scientific disciplines, including real algebraic geometry, proof complexity, control theory, and mathematical programming, and has found applications in fields as diverse as quantum information theory, formal verification, game theory and many others. We survey some connections that were recently uncovered between the Unique Games Conjecture and the Sum-of-Squares method. In particular, we discuss new tools to rigorously bound the running time of the SOS method for obtaining approximate solutions to hard optimization problems, and how these tools give the potential for the sum-of-squares method to provide new guarantees for many problems of interest, and possibly to even refute the UGC.

Mathematics Subject Classification (2010). Primary 68Q25; Secondary 90C22.

Keywords. Sum of squares, semidefinite programming, unique games conjecture, small-set expansion.

1. Introduction

A central mission of theoretical computer science is to understand which computational problems can be solved efficiently, which ones cannot, and what it is about a problem that makes it easy or hard. To illustrate these kind of questions, let us consider the following parameters of an undirected d-regular graph[1] $G = (V, E)$:

- The *smallest connected component of* G is the size of the smallest non-empty set $S \subseteq V$ such that $E(S, V \setminus S) = \emptyset$.
- The *independent-set number of* G is the size of the largest set $S \subseteq V$ such that $E(S, S) = \emptyset$.
- The (*edge*) *expansion*[2] *of* G, denoted ϕ_G, is the minimum *expansion* $\phi_G(S)$ of a vertex

Proceedings of the International Congress of Mathematicians, Seoul, 2014

set $S \subseteq V$ with size $1 \leq |S| \leq |V|/2$, where

$$\phi_G(S) = \frac{|E(S, V \setminus S)|}{d|S|} .$$

The expansion $\phi_G(S)$ measures the probability that a step of the random walk on G leaves S conditioned on starting in S.

All these parameters capture different notions of well-connectedness of the graph G. Computing these can be very useful in many of the settings in which we use graphs to model data, whether it is communication links between servers, social connections between people, genes that are co-expressed together, or transitions between states of a system.

The computational complexity of the first two parameters is fairly well understood. The smallest connected component is easy to compute in time linear in the number $n = |V|$ of vertices by using, for example, breadth-first search from every vertex in the graph. The independent-set number is $\mathbf{NP}$-hard to compute, which means that, assuming the widely believed conjecture that $\mathbf{P} \neq \mathbf{NP}$, it cannot be computed in time polynomial in n. In fact, under stronger (but still widely believed) quantitative versions of the $\mathbf{P} \neq \mathbf{NP}$ conjecture, for every k it is infeasible to decide whether or not the maximum independent set is larger than k in time $n^{o(k)}$ [18, 23] and hence we cannot significantly beat the trivial $O(n^k)$-time algorithm for this problem. Similarly, while we can approximate the independent-set number trivially within a factor of n, assuming such conjectures, there is no polynomial-time algorithm to approximate it within a factor of $n^{1-\varepsilon(n)}$ where $\varepsilon(n)$ is some function tending to zero as n grows [29, 30].

So, connectivity is an easy problem and independent set a hard one, but what about expansion? Here the situation is more complicated. We know that we can't efficiently compute ϕ_G exactly, and we can't even get an arbitrarily good approximation [4], but we actually do have efficient algorithms with non-trivial approximation guarantees for ϕ_G. Discrete versions of *Cheeger's inequality* [2, 3, 17, 21] yield such an estimate, namely

$$\tfrac{d-\lambda_2}{2d} \leq \phi_G \leq 2\sqrt{\tfrac{d-\lambda_2}{2d}} , \tag{1.1}$$

where $\lambda_2(G)$ denotes the (efficiently computable) second largest eigenvalue of the G's adjacency matrix.[3] In particular, we can use (1.1) to efficiently distinguish between graphs with ϕ_G close to 0 and graphs with ϕ_G bounded away from 0. But can we do better? For example, could we efficiently compute a quantity c_G such that $c_G \leq \phi_G \leq O(c_G^{0.51})$? We simply don't know.[4]

[2] An undirected d-regular graph $G = (V, E)$ consists of a set of *vertices* V, which we sometimes identify with the set $[n] = \{1, \ldots, n\}$ for some integer n, and a set of *edges* E, which are 2-element subsets of V, such that every vertex is part of exactly d edges. The assumption that G is regular is not important and made chiefly for notational simplicity. For vertex sets $S, T \subseteq V$, we let $E(S, T)$ denote the set of edges $\{s, t\} \in E$ with $s \in S$ and $t \in T$.

[2] The expansion of a graph is closely related to other quantities, known as *isoperimetric constant*, *conductance* or *sparsest cut*. These quantities are not identical but are the same up to scaling and a multiplicative factor of at most 2. Hence, they are computationally equivalent for our purposes. We also remark that expansion is often not normalized by the degree. However for our purposes this normalization is useful.

[3]The *adjacency matrix* of a graph G is the $|V| \times |V|$ matrix A with 0/1 entries such that $A_{u,v} = 1$ iff $\{u, v\} \in E$.

[4]As we will mention later, there are algorithms to approximate ϕ_G up to factors depending on the number n of vertices, which give better guarantees than (1.1) for graphs where ϕ_G is sufficiently small as a function of n.

This is not an isolated example, but a pattern that keeps repeating. Over the years, computer scientists have developed sophisticated tools to come up with algorithms on one hand, and hardness proofs showing the limits of efficient algorithms on the other hand. But those two rarely match up. Moreover, the cases where we do have tight hardness results are typically in settings, such as the independent set problem, where there is no way to significantly beat the trivial algorithm. In contrast, as a rule, for problems such as computing expansion, where we already know of an algorithm giving non-trivial guarantees, we have no proof that this algorithm is *optimal*. In other words, the following is a common theme:

> *If you already know an algorithm with non-trivial approximation guarantees for a problem, it's very hard to rule out that cleverer algorithms couldn't get even better guarantees.*

In 2002, Subhash Khot formulated a conjecture, known as the *Unique Games Conjecture (UGC)* [31]. A large body of follow up works has shown that this conjecture (whose description is deferred to Section 1.1 below) implies many hardness results that overcome the above challenge and match the best-known algorithms even in cases when they achieve non-trivial guarantees. In fact, beyond just resolving particular questions, this line of works obtained far-reaching complementary *meta algorithmic* and *meta hardness* results. By this we mean results that give an efficient *meta algorithm* $\mathcal{A}$ (i.e., an algorithm that can be applied to a family of problems, and not just a single one) that is *optimal* within a broad domain $\mathcal{C}$, in the sense that (assuming the UGC) there is no polynomial-time algorithm that performs better than $\mathcal{A}$ on any problem in $\mathcal{C}$. It is this aspect of the Unique Games Conjecture result that we find most exciting, and that shows promise of going beyond the current state where the individual algorithmic and hardness results form "isolated islands of knowledge surrounded by a sea of ignorance"[5] into a more unified theory of complexity.

The meta-algorithm that the UGC predicts to be optimal is based on *semidefinite programming* and it uses this technique in a very particular and quite restricted way. (In many settings, this meta-algorithm can be implemented in near-linear time [58].) We will refer to this algorithm as the *UGC meta-algorithm*. It can be viewed as a common generalization of several well known algorithms, including those that underlie Cheeger's Inequality, Grothendieck's Inequality [28], the Goemans–Williamson Max Cut algorithm [24], and the Lovász ϑ function [40]. As we've seen for the example of Cheeger's Inequality, in many of those settings this meta-algorithm gives *non-trivial approximation guarantees* which are the best known, but there are no hardness results ruling out the existence of better algorithms. The works on the UGC has shown that this conjecture (and related ones) imply that this meta-algorithm is *optimal* for a vast number of problems, including all those examples above. For example, a beautiful result of Raghavendra [46] showed that for every constraint-satisfaction problem (a large class of problems that includes many problems of interest such as Max k-SAT, k-Coloring, and Max-Cut), the UGC meta-algorithm gives the best estimate on the maximum possible fraction of constraints one can satisfy. Similarly, the UGC (or closely related variants) imply there are no efficient algorithms that give a better estimate for the sparsest cut of a graph than the one implied by Cheeger's Inequality [51] and no better efficient estimate for the maximum correlation of a matrix with ± 1-valued vectors than the one given by Grothendieck's Inequality.[6] To summarize:

[5] paraphrasing John Wheeler

[6] See [48] for the precise statement of Grothendieck's Inequality and this result. Curiously, the UGC implies that Grothendieck's Inequality yields the best efficient approximation factor for the correlation of a matrix with

> *If true, the Unique Games Conjecture tells us not only* which *problems in a large class are easy and which are hard, but also* why *this is the case. There is a* single unifying reason*, captured by a concrete meta-algorithm, that explains all the easy problem in this class. Moreover, in many cases where this meta-algorithm already gives non-trivial guarantees, the UGC implies that no further efficient improvements are possible.*

All this means that the Unique Games Conjecture is certainly a very attractive proposition, but the big question still remains unanswered—is this conjecture actually true? While some initial results supported the UGC, more recent works, although still falling short of disproving the conjecture, have called it into question. In this survey we discuss the most promising current approach to refute the UGC, which is based on the *Sum of Squares (SOS) method* [37, 42, 45, 55]. The SOS method could potentially refute the Unique Games Conjecture by beating the guarantees of the UGC meta-algorithm on problems on which the conjecture implies the latter's optimality. This of course is interesting beyond the UGC, as it means we would be able to improve the known guarantees for many problems of interest. Alas, analyzing the guarantees of the SOS method is a very challenging problem, and we still have relatively few tools to do so. However, as we will see, we already know that at least in some contexts, the SOS method can yield better results than what was known before. The SOS method is itself a meta algorithm, so even if it turns out to refute the UGC, this does not mean we need to give up on the notion of explaining the complexity of wide swaths of problems via a single algorithm; we may just need to consider a different algorithm. To summarize, regardless of whether it refutes the UGC or not, understanding the power of the SOS method is an exciting research direction that could advance us further towards the goal of a unified understanding of computational complexity.

1.1. The UGC and SSEH conjectures. Instead of the Unique Games Conjecture, in this survey we focus on a related conjecture known as the *Small-Set Expansion Hypothesis (SSEH)* [49]. The SSEH implies the UGC [49], and while there is no known implication in the other direction, there are several results suggesting that these two conjectures are probably equivalent [5, 12, 47, 49, 50]. At any rate, most (though not all) of what we say in this survey applies equally well to both conjectures, but the SSEH is, at least in our minds, a somewhat more natural and simpler-to-state conjecture.

Recall that for a d-regular graph $G = (V, E)$ and a vertex set $S \subseteq V$, we defined its expansion as $\phi_G(S) = |E(S, V \setminus S)|/(d|S|)$. By Cheeger's inequality (1.1), the second largest eigenvalue yields a non-trivial approximation for the minimum expansion $\phi_G = \min_{1 \leq |S| \leq |V|/2} \phi_G(S)$, but it turns out that eigenvalues and similar methods do not work well for the problem of approximating the minimum expansion of smaller sets. The Small-Set Expansion Hypothesis conjectures that this problem is inherently difficult.

Conjecture 1.1 (Small-Set Expansion Hypothesis [49]). *For every $\varepsilon > 0$ there exists $\delta > 0$ such that given any graph $G = (V, E)$, it is* **NP***-hard to distinguish between the case* ***(i)*** *that there exists a subset $S \subseteq V$ with $|S| = \delta|V|$ such that $\phi_G(S) \leq \varepsilon$ and the case* ***(ii)*** *that $\phi_G(S) \geq 1 - \varepsilon$ for every S with $|S| \leq \delta|V|$.*

As mentioned above, the SSEH implies that (1.1) yields an optimal approximation for ϕ_G.

± 1-valued vectors even though we don't actually know the numerical value of this factor (known as Grothendieck's constant).

More formally, assuming the SSEH, there is some absolute constant $c > 0$ such that for every $\phi \geq 0$, it is **NP**-hard to distinguish between the case that a given graph G satisfies $\phi_G \leq \phi$ and the case that $\phi_G \geq c\sqrt{\phi}$ [51]. Given that the SSEH conjectures the difficulty of approximating expansion, the reader might not be so impressed that it also implies the optimality of Cheeger's Inequality. However, we should note that the SSEH merely conjectures that the problem becomes harder as δ becomes smaller, without postulating any quantitative relation between δ and ε, and so it is actually surprising (and requires a highly non-trivial proof) that it implies such quantitatively tight bounds. Even more surprising is that (through its connection with the UGC) the SSEH implies tight hardness result for a host of other problems, including every constraint satisfaction problem, Grothendieck's problem, and many others, which a priori seem to have nothing to do with graph expansion.

Remark 1.2. While we will stick to the SSEH in this survey, for completeness we present here the definition of the Unique Games Conjecture. We will not use this definition in the proceeding and so the reader can feel free to skip this remark. The UGC can be thought of as a more structured variant of the SSEH where we restrict to graphs and sets that satisfy some particular properties. Because we restrict both the graphs and the sets, a priori it is not clear which of these conjectures should be stronger. However it turns out that the SSEH implies the UGC [49]. It is an open problem whether the two conjectures are equivalent, though the authors personally suspect that this is the case.

We say that an n-vertex graph $G = (V, E)$ is δ-*structured* if there is a partition of V into δn sets $V_1, \ldots, V_{\delta n}$ each of size $1/\delta$, such that for every $i \neq j$, either $E(V_i, V_j) = \emptyset$ or $E(V_i, V_j)$ is a *matching* (namely for every $u \in V_i$ there is exactly one $v \in V_j$ such that $\{u, v\} \in E$). We say a set $S \subseteq V$ is δ-*structured* if $|S \cap V_i| = 1$ for all i (and so in particular, $|S| = \delta n$). The Unique Games Conjecture states that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that it is **NP** hard, given a δ-structured G, to distinguish between the case **(i)** that there exists a δ-structured S such that $\phi_G(S) \leq \varepsilon$ and the case **(ii)** that every δ-structured S satisfies $\phi_G(S) \geq 1 - \varepsilon$. The conjecture can also be described in the form of so-called "two prover one round games" (hence its name); see Khot's surveys [32, 33].

1.2. Organization of this survey and further reading. In the rest of this survey we describe the Sum of squares algorithm, some of its applications, and its relation to the Unique Games and Small-Set Expansion Conjectures. We start by defining the Sum of Squares algorithm, and how it relates to classical questions such as Hilbert 17^{th} problem. We will demonstrate how the SOS algorithm is used, and its connection to the UGC/SSEH, by presenting Cheeger's Inequality (1.1) as an instance of this algorithm. The SSEH implies that the SOS algorithm cannot yield better estimates to ϕ_G than those obtained by (1.1). While we do not know yet whether this is true or false, we present two different applications where the SOS does beat prior works— finding a planted sparse vector in a random subspace, and *sparse coding*— learning a set of vectors A given samples of random sparse linear combinations of vectors in A. We then discuss some of the evidence for the UGC/SSEH, how this evidence is challenged by the SOS algorithm and the relation between the UGC/SSEH and the problem of (approximately) finding sparse vectors in arbitrary (not necessarily random) subspaces. Much of our discussion is based on the papers [5, 12–15]. See also [9–11] for informal overviews of some of these issues.

For the reader interested in learning more about the Unique Games Conjecture, there are three excellent surveys on this topic. Khot's CCC survey [33] gives a fairly comprehensive

overview of the state of knowledge on the UGC circa 2010, while his ICM survey [32] focuses on some of the techniques and connections that arose in the works around the UGC. Trevisan [59] gives a wonderfully accessible introduction to the UGC, using the MAX-CUT problem as a running example to explain in detail the UGC's connection to semidefinite programming. As a sign of how rapidly research in this area is progressing, this survey is almost entirely disjoint from [32, 33, 59]. While the former surveys mostly described the implications of the UGC for obtaining very strong hardness and "meta hardness" results, the current manuscript is focused on the question of whether the UGC is actually true, and more generally understanding the power of the SOS algorithm to go beyond the basic LP and SDP relaxations.

Our description of the SOS algorithm barely scratches the surface of this fascinating topic, which has a great many applications that have nothing to do with the UGC or even approximation algorithms at large. The volume [16] and the monograph [39] are good sources for some of these topics. The SOS algorithm was developed in slightly different forms by several researchers, including Shor [55], Nesterov [42], Parrilo [45], and Lasserre [37]. It can be viewed as a strengthening of other "meta-algorithms" proposed by [41, 54] (also known as linear and semi-definite programming hierarchies).[7] Our description of the SOS meta algorithm follows Parrilo's, while the description of the dual algorithm follows Lasserre, although we use the pseudoexpectation notation introduced in [12] instead of Lasserre's notion of "moment matrices". The Positivstellensatz/SOS proof system was first studied by Grigoriev and Vorobjov [27] and Grigoriev [26] proved some degree lower bounds for it, that were later rediscovered and expanded upon by [53, 60]. All these are motivated by the works in real geometry related to Hilbert's 17^{th} problem; see Reznick's survey [52] for more on this research area. One difference between our focus here and much of the other literature on the SOS algorithm is that we are content with proving that the algorithm supplies an *approximation* to the true quantity, rather than exact convergence, but on the other hand are much more stringent about using only very low degree (preferably constant or polylogarithmic in the number of variables).

2. Sums of squares proofs and algorithms

One of the most common ways of proving that a quantity is non-negative is by expressing it as a *Sum of Squares* (SOS). For example, we can prove the Arithmetic-Mean Geometric-Mean inequality $ab \leq a^2/2 + b^2/2$ by the identity $a^2 + b^2 - 2ab = (a-b)^2$. Thus a natural question, raised in the late 19^{th} century, was whether *any* non-negative (possibly multivariate) polynomial can be written as a sum of squares of polynomials. This was answered negatively by Hilbert in 1888, who went on to ask as his 17^{th} problem whether any such polynomial can be written as a sum of squares of *rational* functions. A positive answer was given by Artin [8], and considerably strengthened by Krivine and Stengle. In particular, the following theorem is a corollary of their results, which captures much of the general case.

Theorem 2.1 (Corollary of the Positivstellensatz [36, 57]). *Let $P_1, \ldots, P_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_n]$ be multivariate polynomials. Then, the system of polynomials equations $\mathcal{E} = \{P_1 = 0, \ldots, P_m = 0\}$ has no solution over $\mathbb{R}^n$ if and only if, there exists polynomials*

[7] See [38] for a comparison.

$Q_1, \ldots, Q_m \in \mathbb{R}[x]$ *such that* $S \in \mathbb{R}[x]$ *is a* sum of squares *of polynomials and*

$$-1 = S + \sum Q_i \cdot P_i \,. \tag{2.1}$$

We say that the polynomials $S, Q_1, \ldots, Q_m$ in the conclusion of the theorem form an *SOS proof* refuting the system of polynomial equations[8] $\mathcal{E}$. Clearly the existence of such polynomials implies that $\mathcal{E}$ is unsatisfiable—the interesting part of Theorem 2.1 is the other direction. We say that a SOS refutation $S_1, Q_1, ..., Q_m$ has *degree* ℓ if the maximum degree of the polynomials $Q_i P_i$ involved in the proof is at most ℓ [27]. By writing down the coefficients of these polynomials, we see that a degree-ℓ SOS proof can be written using $mn^{O(\ell)}$ numbers.[9]

In the following lemma, we will prove a special case of Theorem 2.1, where the solution set of $\mathcal{E}$ is a subset of the hypercube $\{\pm 1\}^n$. Here, the degree of SOS refutations is bounded by $2n$. (This bound is not meaningful computationally because the size of degree-$\Omega(n)$ refutations is comparable to the number of points in $\{\pm 1\}$.)

Lemma 2.2. *Let* $\mathcal{E} = \{P_0 = 0, x_1^2 - 1 = 0, \ldots, x_n^2 - 1 = 0\}$ *for some* $P_0 \in \mathbb{R}[x]$. *Then, either the system* $\mathcal{E}$ *is satisfiable or it has a degree-*$2n$ *SOS refutation.*

Proof. Suppose the system is not satisfiable, which means that $P_0(x) \neq 0$ for all $x \in \{\pm 1\}^n$ Since $\{\pm 1\}^n$ is a finite set, we may assume $P_0^2 \geq 1$ over $\{\pm 1\}^n$. Now interpolate the real-valued function $\sqrt{P_0^2 - 1}$ on $\{\pm 1\}^2$ as a multilinear polynomial $R \in \mathbb{R}[x]$. Then, $P_0^2 - 1 - R^2$ is a polynomial of degree at most $2n$ that vanishes over $\{\pm 1\}$, which means that we can write it in the form $\sum_{i=1}^n Q_i \cdot (x_i^2 - 1)$ for polynomials Q_i with $Q_i \leq \deg 2n - 2$. (This fact can be verified either directly or by using that $x_1^2 - 1, \ldots, x_n^2 - 1$ is a Gröbner basis for $\{\pm 1\}^n$.) Putting things together, we see that $-1 = R^2 + (-P_0) \cdot P_0 + \sum_{i=1}^n Q_i \cdot (x_i^2 - 1)$, which is a SOS refutation for $\mathcal{E}$ of the form in Theorem 2.1. □

2.1. From proofs to algorithms. The Sum of Squares algorithm is based on the following theorem, which was discovered in different forms by several researchers:

Theorem 2.3 (SOS Theorem [37, 42, 45, 55], informally stated). *If there is a degree-*ℓ *SOS proof refuting* $\mathcal{E} = \{P_1 = 0, \ldots, P_m = 0\}$*, then such a proof can be found in* $mn^{O(\ell)}$ *time.*

Proof sketch. We can view a degree-ℓ SOS refutation $-1 = S + \sum_i Q_i P_i$ for $\mathcal{E}$ as a system of linear equations in $mn^{O(\ell)}$ variables corresponding to the coefficients of the unknown polynomials $S, Q_1, \ldots, Q_m$. We only need to incorporate the non-linear constraint that S is a sum of squares. But it is not hard to see that a degree-ℓ polynomial S is a sum of squares if and only if there exists a positive-semidefinite matrix M such that $S = \sum_{\alpha,\alpha'} M_{\alpha,\alpha'} x^\alpha x^{\alpha'}$, where α and α' range over all monomials x^α and $x^{\alpha'}$ of degree at most $\ell/2$. Thus, the

[8] In this survey we restrict attention to polynomial *equalities* as opposed to *inequalities*, which turns out to be without loss of generality for our purposes. If we have a system of polynomial inequalities $\{P_1 \geq 0, \ldots, P_m \geq 0\}$ for $P_i \in \mathbb{R}[x]$, the Positivstellensatz certificates of infeasibility take the form $-1 = \sum_{\alpha \subseteq [n]} Q_\alpha P_\alpha$, where each $Q_\alpha \in \mathbb{R}[X]$ is a sum of squares and $P_\alpha = \prod_{i \in \alpha} P_i$. However, we can transform inequalities $\{P_i \geq 0\}$ to equivalent equalities $\{P_i' = P_i - y_i^2 = 0\}$, where $y_1, \ldots, y_m$ are fresh variables. This transformation makes it only easier to find certificates, because $\sum_{\alpha \subseteq [n]} Q_\alpha P_\alpha = S' + \sum_i Q_i' P_i'$ for $S' = \sum_{\alpha \subseteq [n]} Q_\alpha y_\alpha^2$, where $y_\alpha = \prod_{i \in \alpha} y_i$. It also follows that the transformation can only reduce the degree of SOS refutations.

[9] It can be shown that the decomposition of S into sums of squares will not require more than n^ℓ terms; also in all the settings we consider, there are no issues of accuracy in representing real numbers, and so a degree ℓ-proof can be written down using $mn^{O(\ell)}$ bits.

task of finding a degree-ℓ SOS refutation reduces to the task of solving linear systems of equations with the additional constraint that matrix formed by some of the variables is positive-semidefinite. *Semidefinite programming* solves precisely this task and is computationally efficient.[10] □

Remark 2.4 (*What does "efficient" mean?*). In the applications we are interested in, the number of variables n corresponds to our "input size". The equation systems $\mathcal{E}$ we consider can always be solved via a "brute force" algorithm running in $\exp(O(n))$ time, and so degree-ℓ SOS proofs become interesting when ℓ is much smaller than n. Ideally we would want $\ell = O(1)$, though $\ell = \mathrm{polylog}(n)$ or even, say, $\ell = \sqrt{n}$, is still interesting.

Theorem 2.3 yields the following *meta algorithm* that can be applied on any problem of the form

$$\min_{x \in \mathbb{R}^n \,:\, P_1(x) = \cdots = P_m(x) = 0} P_0(x) \tag{2.2}$$

where $P_0, P_1, \ldots, P_m \in \mathbb{R}[x]$ are polynomials. The algorithm is parameterized by a number ℓ called its *degree* and operates as follows:

The degree-ℓ Sum-of-Squares Algorithm

Input: Polynomials $P_0, \ldots, P_m \in \mathbb{R}[x]$

Goal: Estimate $\min P_0(x)$ over all $x \in \mathbb{R}^n$ such that $P_1(x) = \ldots = P_m(x) = 0$

Operation: Output the smallest value $\varphi^{(\ell)}$ such that there does *not* exist a degree-ℓ SOS proof refuting the system,

$$\{P_0 = \varphi^{(\ell)}, P_1 = 0, \ldots, P_m(x) = 0\}\,.^{11}$$

We call $\varphi^{(\ell)}$ the *degree-ℓ SOS estimate* for (2.2), and by Theorem 2.3 it can be computed in $n^{O(\ell)}$ time. For the actual minimum value φ of (2.2), the corresponding system of equations $\{P_0 = \varphi, P_1 = 0, \ldots, P_m = 0\}$ is satisfiable, and hence in particular cannot be refuted by an SOS proof. Thus, $\varphi^{(\ell)} \leq \varphi$ for any ℓ. Since higher degree proofs are more powerful (in the sense that they can refute more equations), it holds that

$$\varphi^{(2)} \leq \varphi^{(4)} \leq \varphi^{(6)} \leq \cdots \leq \min_{x \in \mathbb{R}^n \,:\, P_1(x) = \cdots = P_m(x) = 0} P_0(x)\,.$$

(We can assume degrees of SOS proofs to be even.) As we've seen in Lemma 2.2, for the typical domains we are interested in Computer Science, such as when the set of solutions of $\{P_1 = 0, \ldots, P_m = 0\}$ is equal to $\{\pm 1\}^n$, this sequence is finite in the sense that $\varphi^{(2n)} = \min_{x \in \{\pm 1\}^n} P_0(x)$.

The SOS algorithm uses semidefinite programming in a much more general way than many previous algorithms such as [24, 40]. In fact, the UGC meta-algorithm is the same as the base case (i.e., $\ell = 2$) of the SOS algorithm.

[10] In this survey we ignore issues of numerical accuracy which turn out to be easily handled in our setting.

[11] As in other cases, we are ignoring here issues of numerical accuracy. Also, we note that when actually executing this algorithm, we will not need to check all the (uncountably many) values $\varphi^{(\ell)} \in \mathbb{R}$, but it suffices to enumerate over a sufficiently fine discretization of the interval $[-M, +M]$ for some number M depending on the polynomials $P_0, \ldots, P_m$. This step can be carried out in polynomial time in all the settings we consider.

Recall that the UGC and SSEH imply that in many settings, one cannot improve on the approximation guarantees of the UGC meta-algorithm without using $\exp(n^{\Omega(1)})$ time. Thus in particular, if those conjectures are true then in those settings, using the SOS meta algorithm with degree, say, $\ell = 10$ (or even $\ell = \mathrm{polylog}(n)$ or $\ell = n^{o(1)}$) will not yield significantly better guarantees than $\ell = 2$.

Remark 2.5 (Comparison with local-search based algorithms). Another approach to optimize over non-linear problems such as (2.2) is to use local-search algorithms such as gradient descent that make local improvement steps, e.g., in the direction of the gradient, until a local optimum is reached. One difference between such local search algorithms and the SOS algorithm is that the latter sometimes succeeds in optimizing highly non-convex problems that have exponential number of local optima. As an illustration, consider the polynomial $P(x) = n^4 \sum_{i=1}^n (x_i^2 - x_i)^2 + (\sum_{i=1}^n x_i)^2$.

Its unique global minimum is the point $x = 0$, but it is not hard to see that it has an exponential number of local minima (for every $x \in \{0,1\}^n$, $P(x) < P(y)$ for every y with $\|y - x\| \in [1/n, 2/n]$, and so there must be a local minima in the ball of radius $1/n$ around x). Hence, gradient descent or other such algorithms are extremely likely to get stuck in one of these suboptimal local minima. However, since P is in fact a sum of squares with constant term 0, the degree-4 SOS algorithm will output P's correct global minimum value.

2.2. Pseudodistributions and pseudoexpectations. Suppose we want to show that the level-ℓ SOS meta-algorithm achieves a good approximation of the minimum value of P_0 over the set $\mathcal{Z} = \{x \in \mathbb{R}^n \mid P_1(x) = \cdots = P_m(x) = 0\}$ for a particular kind of polynomials $P_0, P_1, \ldots, P_m \in \mathbb{R}[x]$. Since the estimate $\varphi^{(\ell)}$ always lower bounds this quantity, we are to show that

$$\min_{\mathcal{Z}} P_0 \leq f(\varphi^{(\ell)}) \tag{2.3}$$

for some particular function f (satisfying $f(\varphi) \geq \varphi$) which captures our approximation guarantee. (E.g., a factor c approximation corresponds to the function $f(\varphi) = c\varphi$.)

If we expand out the definition of $\varphi^{(\ell)}$, we see that to prove Equation (2.3) we need to show that for every φ if there does not exists a degree-ℓ proof that $P_0(x) \neq \varphi$ for all $x \in \mathcal{Z}$, then there exists an $x \in \mathcal{Z}$ such that $P_0(x) \leq f(\varphi)$. So, to prove a result of this form, we need to find ways to use the *non-existence* of a proof. Here, *duality* is useful.

> *Pseudodistributions are the dual object to SOS refutations, and hence the* non-existence *of a refutation implies the* existence *of a pseudodistribution.*

We now elaborate on this, and explain both the definition and intuition behind pseudodistributions. In Section 3 we will give a concrete example, by showing how one can prove that degree-2 SOS proofs capture Cheeger's Inequality using such an argument. Results such as the analysis of the Goemans-Williamson Max Cut algorithm [24], and the proof of Grothendieck's Inequality [28] can be derived using similar methods.

Definition 2.6. Let $\mathbb{R}[x]_\ell$ denote the set of polynomials in $\mathbb{R}[x]$ of degree at most ℓ. A *degree-ℓ pseudoexpectation operator for* $\mathbb{R}[x]$ is a linear operator $\mathcal{L}$ that maps polynomials in $\mathbb{R}[x]_\ell$ into $\mathbb{R}$ and satisfies that $\mathcal{L}(1) = 1$ and $\mathcal{L}(P^2) \geq 0$ for every polynomial P of degree at most $\ell/2$.

The term pseudoexpectation stems from the fact that for every distribution $\mathcal{D}$ over $\mathbb{R}^n$, we can obtain such an operator by choosing $\mathcal{L}(P) = \mathbb{E}_{\mathcal{D}} P$ for all $P \in \mathbb{R}[x]$. Moreover, the

properties $\mathcal{L}(1) = 1$ and $\mathcal{L}(P^2) \geq 0$ turn out to capture to a surprising extent the properties of distributions and their expectations that we tend to use in proofs. Therefore, we will use a notation and terminology for such pseudoexpectation operators that parallels the notation we use for distributions. In fact, all of our notation can be understood by making the thought experiment that there exists a distribution as above and expressing all quantities in terms of low-degree moments of that distribution (so that they also make sense if we only have a pseudoexpectation operator that doesn't necessarily correspond to a distribution).

In the following, we present the formal definition of our notation. We denote pseudoexpectation operators as $\tilde{\mathbb{E}}_{\mathcal{D}}$, where $\mathcal{D}$ acts as index to distinguish different operators. If $\tilde{\mathbb{E}}_{\mathcal{D}}$ is a degree-ℓ pseudoexpectation operator for $\mathbb{R}[x]$, we say that $\mathcal{D}$ is a *degree-ℓ pseudodistribution* for the indeterminates x. In order to emphasize or change indeterminates, we use the notation $\tilde{\mathbb{E}}_{y\sim\mathcal{D}} P(y)$. In case we have only one pseudodistribution $\mathcal{D}$ for indeterminates x, we denote it by $\{x\}$. In that case, we also often drop the subscript for the pseudoexpectation and write $\tilde{\mathbb{E}} P$ for $\tilde{\mathbb{E}}_{\{x\}} P$.

We say that a degree-ℓ pseudodistribution $\{x\}$ satisfies a system of polynomial equations $\{P_1 = 0, \ldots, P_m = 0\}$ if $\tilde{\mathbb{E}} Q \cdot P_i = 0$ for all $i \in [m]$ and all polynomials $Q \in \mathbb{R}[x]$ with $\deg Q \cdot P_i \leq \ell$. We also say that $\{x\}$ satisfies the constraint $\{P(x) \geq 0\}$ if there exists some sum-of-squares polynomial $S \in \mathbb{R}[x]$ such that $\{x\}$ satisfies the polynomial equation $\{P = S\}$. It is not hard to see that if $\{x\}$ was an actual distribution, then these definitions imply that all points in the support of the distribution satisfy the constraints. We write $P \succcurlyeq 0$ to denote that P is a sum of squares of polynomials, and similarly we write $P \succcurlyeq Q$ to denote $P - Q \succcurlyeq 0$.

The duality between SOS proofs and pseudoexpectations is expressed in the following theorem. We say that a system $\mathcal{E}$ of polynomial equations is *explicitly bounded* if there exists a linear combination of the constraints in $\mathcal{E}$ that has the form $\{\sum_i x_i^2 + S = M\}$ for $M \in \mathbb{R}$ and $S \in \mathbb{R}[x]$ a sum-of-squares polynomial. (Note that in this case, every solution $x \in \mathbb{R}^n$ of the system $\mathcal{E}$ satisfies $\sum_i x_i^2 \leq M$.)

Theorem 2.7. *Let $\mathcal{E} = \{P_1 = 0, \ldots, P_m = 0\}$ be a set of polynomial equations with $P_i \in \mathbb{R}[x]$. Assume that $\mathcal{E}$ is explicitly bounded in the sense above. Then, exactly one of the following two statements holds: (a) there exists a degree-ℓ SOS proof refuting $\mathcal{E}$, or (b) there exists a degree-ℓ pseudodistribution $\{x\}$ that satisfies $\mathcal{E}$.*

Proof. First, suppose there exists a degree-ℓ refutation of the system $\mathcal{E}$, i.e., there exists polynomials $Q_1, \ldots, Q_m \in \mathbb{R}[x]$ and a sum-of-squares polynomial $R \in \mathbb{R}[x]$ so that $-1 = R + \sum_i Q_i P_i$ and $\deg Q_i P_i \leq \ell$. Let $\{x\}$ be any pseudodistribution. We are to show that $\{x\}$ does not satisfy $\mathcal{E}$. Indeed, $\tilde{\mathbb{E}} \sum_i Q_i P_i = -\tilde{\mathbb{E}} 1 - \tilde{\mathbb{E}} R \leq -1$, which means that $\tilde{\mathbb{E}} Q_i P_i \neq 0$ for at least one $i \in [m]$. Therefore, $\{x\}$ does not satisfy $\mathcal{E}$.

Next, suppose there does not exist a degree-ℓ refutation of the system $\mathcal{E}$. We are to show that there exists a pseudodistribution that satisfies $\mathcal{E}$. Let $\mathcal{C}$ be the cone of all polynomials of the form $R + \sum_i Q_i P_i$ for sum-of-squares R and polynomials Q_i with $\deg Q_i P_i \leq \ell$. Since $\mathcal{E}$ does not have a degree-ℓ refutation, the constant polynomial -1 is not contained in $\mathcal{C}$. We claim that from our assumption that the system $\mathcal{E}$ is explicitly bounded it follows that -1 also cannot lie on the boundary of $\mathcal{C}$. Assuming this claim, the hyperplane separation theorem implies that there exists a linear form L such that $L(-1) < 0$ but $L(P) \geq 0$ for all $P \in \mathcal{C}$. By rescaling, we may assume that $L(1) = 1$. Now this linear form satisfies all conditions of a pseudoexpectation operator for the system $\mathcal{E}$.

Proof of claim. We will show that if -1 lies on the boundary of $\mathcal{C}$, then also $-1 \in \mathcal{C}$. If -1 is on the boundary of $\mathcal{C}$, then there exists a polynomial $P \in \mathbb{R}[X]_\ell$ such that $-1 + \varepsilon P \in \mathcal{C}$ for all $\varepsilon > 0$ (using the convexity of $\mathcal{C}$). Since $\mathcal{E}$ is explicitly bounded, for every polynomial $P \in \mathbb{R}[X]_\ell$, the cone $\mathcal{C}$ contains a polynomial of form $N - P - R$ for a sum-of-square R and a number N. (Here, the polynomial $N - P - R \in \mathcal{C}$ is a certificate that $P \leq N$ over the solution set of $\mathcal{E}$. Such a certificate is easy to obtain when $\mathcal{E}$ is explicitly bounded. We are omitting the details.) At this point, we see that -1 is a nonnegative combination of the polynomials $-1 + \varepsilon P$, $N - P - R$, and R for $\varepsilon < 1/N$. Since these polynomials are contained in $\mathcal{C}$, their nonnegative combination -1 is also contained in the cone $\mathcal{C}$. □

Recipe for using pseudoexpectations algorithmically. In many applications we will use the following dual form of the SOS algorithm:

The degree-ℓ Sum-of-Squares Algorithm (dual form)
Input: Polynomials $P_0, \ldots, P_m \in \mathbb{R}[x]$
Goal: Estimate $\min P_0(x)$ over all x with $P_1(x) = \ldots = P_m(x) = 0$
Operation: Output the smallest value $\varphi^{(\ell)}$ such that there is a degree-ℓ pseudodistribution $\{x\}$ satisfying the system,

$$\{P_0 = \varphi^{(\ell)}, P_1 = 0, \ldots, P_m(x) = 0\}.$$

Theorem 2.7 shows that in the cases we are interested in, both variants of the SOS algorithm will output the same answer. Regardless, a similar proof to that of Theorem 2.3 shows that the dual form of the SOS algorithm can also be computed in time $n^{O(\ell)}$. Thus, when using the SOS meta-algorithm, instead of trying to argue from the non-existence of a proof, we will use the existence of a pseudodistribution. Specifically, to show that the algorithm provides an $f(\cdot)$ approximation in the sense of (2.3), what we need to show is that given a degree-ℓ pseudodistribution $\{x\}$ satisfying the system $\{P = \varphi, P_1 = 0, \ldots, P_m = 0\}$, we can find some particular x^* that satisfies $P(x^*) \leq f(\varphi)$. Our approach to doing so (based on the authors' paper with Kelner [15]) can be summarized as follows:

> *Solve the problem pretending that $\{x\}$ is an* actual distribution *over solutions, and if all the steps you used have low-degree SOS proofs, the solution still works even when $\{x\}$ is a low-degree* pseudodistribution.

It may seem that coming up with an algorithm for the actual distribution case is trivial, as any element in the support of the distribution would be a good solution. However note that even in the case of a real distribution, the algorithm does not get sampling access to the distribution, but only access to its low-degree moments. Depending on the reader's temperament, the above description of the algorithm, which "pretends" pseudodistributions are real ones, may sound tautological or just wrong. Hopefully it will be clearer after the next two sections, where we use this approach to show how the SOS algorithm can match the guarantee of Cheeger's Inequality for computing the expansion, to find planted sparse vectors in random subspaces, and to approximately recover sparsely used dictionaries.

3. Approximating expansion via sums of squares

Recall that the *expansion*, ϕ_G, of a d-regular graph $G = (V, E)$ is the minimum of $\phi_G(S) = |E(S, V \setminus S)||/(d|S|)$ over all sets S of size at most $|V|/2$. Letting $x = \mathbb{1}_S$ be the characteristic vector[12] of the set S the expression $|E(S, V \setminus S)|$ can be written as $\sum_{\{i,j\}\in E}(x_i - x_j)^2$ which is a quadratic polynomial in x. Therefore, for every k, computing the value $\phi_G(k) = \min_{|S|=k} |E(S, V \setminus S)|/(dk)$ can be phrased as the question of minimizing a polynomial P_0 over the set of x's satisfying the equations $\{x_i^2 - x_i = 0\}_{i=1}^n$ and $\{\sum_{i=1}^n x_i = k\}$. Let $\phi_G^{(\ell)}(k)$ be the degree-ℓ SOS estimate for $\phi_G(k)$. We call $\phi_G^{(\ell)} = \min_{k\leq n/2} \phi_G(k)$ the degree-ℓ SOS estimate for ϕ_G. Note that $\phi_G^{(\ell)}$ can be computed in $n^{O(\ell)}$ time. For the case $\ell = 2$, the following theorem describes the approximation guarantee of the estimate $\phi_G^{(\ell)}$.

Theorem 3.1. *There exists an absolute constant c such that for every graph G*

$$\phi_G \leq c\sqrt{\phi_G^{(2)}} \tag{3.1}$$

Before we prove Theorem 3.1, let us discuss its significance. Theorem 3.1 is essentially a restatement of Cheeger's Inequality in the SOS language—the degree 2-SOS algorithm is the UGC meta algorithm which is essentially the same as the algorithm based on the second-largest eigenvalue.[13] There are examples showing that (3.1) is tight, and so we cannot get better approximation using degree 2 proofs. But can we get a better estimate using degree 4 proofs? Or degree $\log n$ proofs? We don't know the answer, but if the Small-Set Expansion Hypothesis is true, then beating the estimate (3.1) is **NP** -hard, which means (under standard assumptions) that to do so we will need to use proofs of degree at least $n^{\Omega(1)}$.

This phenomenon repeats itself in other problems as well. For example, for both the Grothendieck Inequality and the Max Cut problems, the SSEH (via the UGC) predicts that beating the estimate obtained by degree-2 proofs will require degree $\ell = n^{\Omega(1)}$. As in the case of expansion, we have not been able to confirm or refute these predictions. However, we will see some examples where using higher degree proofs *does* help, some of them suspiciously close in nature to the expansion problem.

One such example comes from the beautiful work of Arora, Rao and Vazirani [7] who showed that

$$\phi_G \leq O(\sqrt{\log n}) \cdot \phi_G^{(6)} \ ,$$

which is better than the guarantee of Theorem 3.1 for $\phi_G \ll 1/\log n$. However, this is not known to contradict the SSEH or UGC, which apply to the case when ϕ_G is a small constant.

As we will see in Section 5, for the small set expansion problem of approximating $\phi_G(S)$ for small sets S, we can beat the degree 2 bounds with degree $\ell = n^\tau$ proofs where τ is a parameter tending to zero with the parameter ε of the SSEH [5]. This yields a sub-exponential algorithm for the small-set expansion problem (which can be extended to the Unique Games problem as well) that "barely misses" refuting the SSEH and UGC. We will also see that degree $O(1)$ proofs have surprising power in other settings that are closely related to the SSEH/UGC, but again at the moment still fall short of refuting those conjectures.

[12] The i-th coordinate of vector $\mathbb{1}_S$ is equal 1 if $i \in S$ and equal 0 otherwise.

[13] The second-largest eigenvalue is directly related to the minimum value of φ such that there exists a degree-2 pseudodistribution satisfying the more relaxed system $\{\sum_{\{ij\}\in E}(x_i - x_j)^2 = \varphi \cdot dn/2, \sum_i x_i = n/2, \sum_i x_i^2 = n/2\}$.

3.1. Proof of theorem 3.1. This proof is largely a reformulation of the standard proof of a discrete variant of Cheeger's Inequality, phrased in the SOS language of pseudodistributions, and hence is included here mainly to help clarify these notions, and to introduce a tool—sampling from a distribution matching first two moments of a pseudodistribution— that will be useful for us later on. By the dual formulation, to prove Theorem 3.1 we need to show that given a pseudodistribution $\{x\}$ over characteristic vectors of size-k sets S of size $k \leq n/2$ with $|E(S, V \setminus S)| = \varphi dk$, we can find a particular set S^* of size at most $n/2$ such that $E(S^*, V \setminus S^*) \leq O(\sqrt{\varphi})d|S^*|$. For simplicity, we consider the case $k = n/2$ (the other cases can be proven in a very similar way). The distribution $\{x\}$ satisfies the constraints $\{\sum x_i = n/2\}$, $\{x_i^2 = x_i\}$ for all i, and $\{\sum_{\{i,j\}\in E}(x_i - x_j)^2 = \varphi d \sum_i x_i\}$. The algorithm to find S^* is quite simple:

1. Choose $(y_1, \ldots, y_n)$ from a random Gaussian distribution with the same quadratic moments as $\{x\}$ so that $\mathbb{E}\, y_i = \tilde{\mathbb{E}} x_i$ and $\mathbb{E}\, y_i y_j = \tilde{\mathbb{E}} x_i x_j$ for all $i, j \in [n]$. (See details below.)
2. Output the set $S^* = \{i \mid y_i \geq 1/2\}$ (which corresponds to the 0/1 vector closest to y).

We remark that the set produces by the algorithm might have cardinality larger than $n/2$, in which case we will take the complement of S^*.

Sampling from a distribution matching two moments. We will first give a constructive proof the well-known fact that for every distribution over $\mathbb{R}^n$, there exists an n-dimensional Gaussian distribution with the same quadratic moments. Given the moments of a distribution $\{x\}$ over $\mathbb{R}^n$, we can sample a Gaussian distribution $\{y\}$ matching the first two moments of $\{x\}$ as follows. First, we can assume $\mathbb{E}\, x_i = 0$ for all i by shifting variables if necessary. Next, let $v^1, \ldots, v^n$ and $\lambda_1, \ldots, \lambda_n$ be the eigenvectors and eigenvalues of the matrix $M_{i,j} = \mathbb{E}\, x_i x_j$. (Note that M is positive semidefinite and so $\lambda_1, \ldots, \lambda_n \geq 0$.) Choose i.i.d random standard Gaussian variables $w_1, \ldots, w_n$ and define $y = \sum_k \sqrt{\lambda_k} w_k v^k$. Since $\mathbb{E}\, w_k w_{k'}$ equals 1 if $k = k'$ and equals 0 otherwise,

$$\mathbb{E}\, y_i y_j = \sum_k \lambda_k (v^k)_i (v^k)_j = M_{i,j} \ .$$

One can verify that if $\{x\}$ is a degree-2 pseudodistribution then the second moment matrix M of the shifted version of x (such that $\tilde{\mathbb{E}} x_i = 0$ for all i) is positive-semidefinite, and hence the above can be carried for pseudodistributions of degree at least 2 as well. Concretely, if we let $\bar{x} = \tilde{\mathbb{E}} x$ be the mean of the pseudodistribution, then $M = \tilde{\mathbb{E}}(x - \bar{x})(x - \bar{x})^\top$. This matrix is positive semidefinite because every test vector $z \in \mathbb{R}^n$ satisfies $z^\top M z = \tilde{\mathbb{E}}\big(z^\top (x - \bar{x})\big)^2 \geq 0$.

Analyzing the algorithm. The analysis is based on the following two claims: (i) the set S^* satisfies $n/3 \leq |S^*| \leq 2n/3$ with constant probability and (ii) in expectation $|E(S^*, V \setminus S^*)| \leq O(\sqrt{\varphi} dn)$.

We will focus on two extreme cases that capture the heart of the arguments for the claims. In the first case, all variables y_i have very small variance so that $\mathbb{E}\, y_i^2 \approx (\mathbb{E}\, y_i)^2$. In this case, because our constraints imply that $\mathbb{E}\, y_i^2 = \mathbb{E}\, y_i$, every variable satisfies either $\mathbb{E}\, y_i^2 \approx 0$ or $\mathbb{E}\, y_i^2 \approx 1$, which means that the distribution of the set S^* produced by the algorithm is concentrated around a particular set, and it is easy to verify that this set satisfies the two

claims. In the second, more interesting case, all variables y_i have large variance, which means $\mathbb{E}\, y_i^2 = 1/2$ in our setting.

In this case, each event $\{y_i \geq 1/2\}$ has probability $1/2$ and therefore $\mathbb{E}|S^*| = n/2$. Using that the quadratic moments of $\{y\}$ satisfy $\mathbb{E}\sum_i y_i = n/2$ and $\mathbb{E}(\sum_i y_i)^2 = (n/2)^2$, one can show that these events cannot be completely correlated, which allows us to control the probability of the event $n/3 \leq |S^*| \leq 2n/3$ and establishes (i). For the second claim, it turns out that by convexity considerations it suffices to analyze the case that all edges contribute equally to the term $\frac{1}{|E|}\sum_{\{i,j\}\in E} \tilde{\mathbb{E}}(x_i - x_j)^2 = \varphi$, so that $\tilde{\mathbb{E}}(x_i - x_j)^2 = \varphi$ for all $\{i,j\} \in E$. So we see that $\{y_i, y_j\}$ is a 2-dimensional Gaussian distribution with mean $(\frac{1}{2}, \frac{1}{2})$ and covariance $\frac{1}{4}\left(\begin{smallmatrix} 1 & 1-2\varphi \\ 1-2\varphi & 1 \end{smallmatrix}\right)$ Thus, in order to bound the expected value of $|E(S^*, V \setminus S^*)|$, we need to bound the probability of the event "$y_i \geq 1/2$ and $y_j < 1/2$" for this particular Gaussian distribution, which amounts to a not-too-difficult calculation that indeed yields an upper bound of $O(\sqrt{\varphi})$ on this probability. □

4. Machine learning with sum of squares

In this section, we illustrate the computational power of the sum-of-squares method with applications to two basic problems in unsupervised learning. In these problems, we are given samples of an unknown distribution from a fixed, parametrized family of distributions and the goal is to recover the unknown parameters from these samples. Despite the average-case nature of these problems, most of the analysis in these applications will be for deterministic problems about polynomials that are interesting in their own right.

The first problem is SPARSE VECTOR RECOVERY. Here, we are given a random basis of a d-dimensional linear subspace $U \subseteq \mathbb{R}^n$ of the form

$$U = \mathrm{Span}\{x^{\langle}0, x^{\langle}1, \ldots, x^{\langle}d\}\,,$$

where $x^{\langle}0$ is a sparse vector and $x^{(1)}, \ldots, x^{\langle}d$ are independent standard Gaussian vectors. The goal is to reconstruct the vector $x^{(0)}$. This is a natural problem in its own right, and is also a useful subroutine in various settings; see [20]. Demanet and Hand [20] gave an algorithm (based on [56]) that recovers $x^{(0)}$ by searching for the vector x in U that maximizes $\|x\|_\infty/\|x\|_1$ (which can be done efficiently by n linear programs). It is not hard to show that $x^{(0)}$ has to have less than $|n|/\sqrt{d}$ coordinates for it to be maximize this ratio,[14] and hence this was a limitation of prior techniques. In contrast, as long as d is not too large (namely, $d = O(\sqrt{n})$), the SOS method can recover $x^{(0)}$ as long as it has less than εn coordinates for some constant $\varepsilon > 0$ [15].

The second problem we consider is SPARSE DICTIONARY LEARNING, also known as SPARSE CODING. Here, we are given independent samples $y^{\langle}1, \ldots, y^{\langle}R \in \mathbb{R}^n$ from an unknown distribution of the form $\{y = Ax\}$, where $A \in \mathbb{R}^{n\times m}$ is a matrix and x is a random m-dimensional vector from a distribution over sparse vectors. This problem, initiated by the work Olshausen and Field [44] in computational neuroscience, has found a variety of uses in machine learning, computer vision, and image processing (see, e.g. [1] and the references therein). The appeal of this problem is that intuitively data should be sparse in the "right" representation (where every coordinate corresponds to a meaningful feature), and finding this

[14] See Lemma 5.2 below for a related statement

representation can be a useful first step for further processing, just as representing sound or image data in the Fourier or Wavelet bases is often a very useful primitive. While there are many heuristics use to solve this problem, prior works giving rigorous recovery guarantees such as [1, 6, 56] all required the vector x to be *very* sparse, namely less than $\sqrt{n}$ nonzero entries.[15] In contrast, the SOS method can be used to approximately recover the dictionary matrix A as long as x has $o(n)$ nonzero (or more generally, significant) entries [14].

4.1. Sparse vector recovery. We say a vector x is μ-sparse if the 0/1 indicator $\mathbb{1}_{\operatorname{supp} x}$ of the support of x has norm-squared $\mu = \|\mathbb{1}_{\operatorname{supp} x}\|_2^2$. The ratio $\mu/\|\mathbb{1}\|_2^2$ is the fraction of non-zero coordinates in x.

Theorem 4.1. *There exists a polynomial-time approximation algorithm for* SPARSE VECTOR RECOVERY *with the following guarantees: Suppose the input of the algorithm is an arbitrary basis of a $d+1$-dimensional linear subspace $U \subseteq \mathbb{R}^n$ of the form $U = \operatorname{Span}\{x^{(0)}, x^{(1)} \dots, x^{()}d\}$ such that $x^{()}0$ is a μ-sparse unit vector with $\mu \le \varepsilon \cdot \|\mathbb{1}\|_2^2$ and $x^{(1)}, \dots, x^{()}d$ are standard Gaussian vectors orthogonal to $x^{()}0$ with $d \ll \sqrt{n}$. Then, with probability close to 1, the algorithm outputs a unit vector x that has correlation $\langle x, x^{()}0\rangle^2 \ge 1 - O(\varepsilon)$ with $x^{(0)}$.*

Our algorithm will follow the general recipe we described in Section 2.2:

> *Find a system of polynomial equations $\mathcal{E}$ that captures the intended solution $x^{()}0$, then pretend you are given a distribution $\{u\}$ over solutions of $\mathcal{E}$ and show how you could recover a single solution u^* from the low order moments of $\{u\}$.*

Specifically, we come up with a system $\mathcal{E}$ so that desired vector $x^{()}0$ satisfies all equations, and it is essentially the only solution to $\mathcal{E}$. Then, using the SOS algorithm, we compute a degree-4 pseudodistribution $\{u\}$ that satisfies $\mathcal{E}$. Finally, as in Section 3.1, we sample a vector u^* from a Gaussian distribution that has the same quadratic moments as the pseudodistribution $\{u\}$.

How to encode this problem as a system of polynomial equations? By Cauchy–Schwarz, any μ-sparse vector x satisfies $\|x\|_2^2 \le \|x\|_{2p}^2 \cdot \|\mathbb{1}_{\operatorname{supp} x}\|_q = \|x\|_{2p}^2 \cdot \mu^{1-1/p}$ for all $p, q \ge 1$ with $1/p + 1/q = 1$. In particular, for $p = 2$, such vectors satisfy $\|x\|_4^4 \ge \|x\|_2^4/\mu$. This fact motivates our encoding of SPARSE VECTOR RECOVERY as a system of polynomial equations. If the input specifies subspace $U \subseteq \mathbb{R}^n$, then we compute the projector P into the subspace U and choose the following polynomial equations: $\|u\|_2^2 = 1$ and $\|Pu\|_4^4 = 1/\mu_0$, where $\mu_0 = \|x_0\|_2^4/\|x_0\|_4^4$. (We assume here the algorithm is given $\mu_0 \le \mu$ as input, as we can always guess a sufficiently close approximation to it.)

Why does the sum-of-squares method work? The analysis of algorithm has two ingredients. The first ingredient is a structural property about projectors of random subspaces.

Lemma 4.2. *Let $U' \subseteq \mathbb{R}^n$ be a random d-dimensional subspace with $d \ll \sqrt{n}$ and let P' be the projector into U'. Then, with high probability, the following sum-of-squares relation over $\mathbb{R}[u]$ holds for $\mu' \ge \Omega(1) \cdot \|\mathbb{1}\|_2^2$,*

$$\|P'u\|_4^4 \preccurlyeq \|u\|_2^4/\mu' .$$

[15] If the distribution x consists of m independent random variables then better guarantees can be achieved using *Independent Component Analysis (ICA)* [19]. See [25] for the current state of art in this setting. However we are interested here in the more general case.

Proof outline. We can write $P' = B^\top B$ where B is a $d \times n$ matrix whose rows are an orthogonal basis for the subspace U'. Therefore, $Pu = B^\top x$ where $x = Bu$, and so to prove Lemma 4.2 it suffices to show that under these conditions, $\|B^\top x\|_4^4 \preccurlyeq O(\|x\|_2^4/\|\mathbb{1}\|_2^4)$. The matrix $B^\top$ will be very close to having random independent Gaussian entries, and hence, up to scaling, $\|B^\top x\|_4^4$ will be (up to scaling), close to $Q(x) = \frac{1}{n}\sum\langle w_i, x\rangle^4$ where $w_1, \ldots, w_d \in \mathbb{R}^d$ are chosen independently at random from the standard Gaussian distribution. The expectation of $\langle w, x\rangle^4$ is equal $3\sum_{i,j} x_i^2 x_j^2 = 3\|x\|_2^4$. Therefore, to prove the lemma, we need to show that for $n \gg d^2$, the polynomial $Q(x)$ is with high probability close to its expectation, in the sense that the $d^2 \times d^2$ matrix corresponding to Q's coefficients is close to its expectation in the spectral norm. This follows from standard matrix concentration inequalities, see [12, Theorem 7.1[16]]). □

The following lemma is the second ingredient of the analysis of the algorithm.

Lemma 4.3. *Let $U' \subseteq \mathbb{R}^n$ be a linear subspace and let P' be the projector into U'. Let $x^\circ 0 \in \mathbb{R}^n$ be a μ-sparse unit vector orthogonal to U' and let $U = \mathrm{Span}\{x^\circ 0\} \oplus U'$ and P the projector on U. Let $\{u\}$ be a degree-4 pseudodistribution that satisfies the constraints $\{\|u\|_2^2 = 1\}$ and $\{\|Pu\|_4^4 = 1/\mu_0\}$, where $\mu_0 = \|x^\circ 0\|_2^4/\|x^\circ 0\|_4^4 \le \mu$. Suppose $\|P'u\|_4^4 \preccurlyeq \|u\|_2^4/\mu'$ is a sum-of-squares relation in $\mathbb{R}[u]$. Then, $\{u\}$ satisfies*

$$\tilde{\mathbb{E}}\|P'u\|_2^2 \le 4\big(\tfrac{\mu}{\mu'}\big)^{1/4}.$$

Note that the conclusion of Lemma 4.3 implies that a vector u^* sampled from a Gaussian distribution with the same quadratic moments as the computed pseudodistribution also satisfies $\mathbb{E}_{u^*}\|P'u^*\|_2^2 \le 4(\mu/\mu')^{1/4}$ and $\mathbb{E}\|u^*\|_2^2 = 1$. By Markov inequality, $\|u^* - x^\circ 0\|_2^2 \le 16(\mu/\mu')^{1/4}$ holds with probability at least $3/4$. Since u^* is Gaussian, it satisfies $\|u^*\|_2^2 \ge 1/4$ with probability at least $1/2$. If both events occur, which happens with probability at least $1/4$, then $\langle u^*, x^\circ 0\rangle^2 \ge (1 - O(\mu/\mu'))\|u^*\|_2^2$, thus establishing Theorem 4.1.

Proof of lemma 4.3 There are many ways in which pseudodistributions behave like actual distributions, as far as low degree polynomials are concerned. To prove Lemma 4.3, we need to establish the following two such results:

Lemma 4.4 (Hölder's inequality for pseudoexpectation norms). *Suppose a and b are non-negative integers that sum to a power of 2. Then, every degree-$(a+b)$ pseudodistribution $\{u, v\}$ satisfies*

$$\tilde{\mathbb{E}}\,\mathbb{E}_i\, u_i^a v_i^b \le \Big(\tilde{\mathbb{E}}\,\mathbb{E}_i\, u_i^{a+b}\Big)^{a/(a+b)} \cdot \Big(\tilde{\mathbb{E}}\,\mathbb{E}_i\, v_i^{a+b}\Big)^{b/(a+b)}.$$

Proof sketch. The proof of the general case follows from the case $a = b = 2$ by an inductive argument. The proof for the case $a = b = 1$ follows from the fact that the polynomial $\alpha\,\mathbb{E}_i\, u_i^2 + \beta\,\mathbb{E}_i\, v_i^2 - \sqrt{\alpha\beta}\,\mathbb{E}_i\, u_i v_i \in \mathbb{R}[u, v]$ is a sum of squares for all $\alpha, \beta \ge 0$ and choosing $\alpha = 1/\tilde{\mathbb{E}}\,\mathbb{E}_i\, u_i^2$ and $\beta = 1/\tilde{\mathbb{E}}\,\mathbb{E}_i\, v_i^2$ □

Lemma 4.5 (Triangle inequality for pseudodistribution ℓ_4 norm). *Let $\{u, v\}$ be a degree-4 pseudodistribution. Then,*

$$\Big(\tilde{\mathbb{E}}\|u + v\|_4^4\Big)^{1/4} \le \Big(\tilde{\mathbb{E}}\|u\|_4^4\Big)^{1/4} + \Big(\tilde{\mathbb{E}}\|v\|_4^4\Big)^{1/4}.$$

[16]The reference is for the arxiv version arXiv:1205.4484v2 of the paper.

Proof. The inequality is invariant with respect to the measure used for the inner norm $\|\cdot\|_4$. For simplicity, suppose $\|x\|_4^4 = \mathbb{E}\, x_i^4$. Then, $\|u+v\|_4^4 = \mathbb{E}_i\, u_i^4 + 4\,\mathbb{E}_i\, u_i^3 v_i + 6\,\mathbb{E}_i\, u_i v_i^3 + \mathbb{E}_i\, v_i^4$. Let $A = \tilde{\mathbb{E}}\,\mathbb{E}_i\, u_i^4$ and $B = \tilde{\mathbb{E}}\,\mathbb{E}_i\, v_i^4$. Then, Lemma 4.5 allows us to bound the pseudoexpectations of the terms $\mathbb{E}_i\, u_i^a v_i^b$, so that as desired

$$\tilde{\mathbb{E}}\|u+v\|_4^4 \le A + 4A^{3/4}B^{1/4} + 6A^{1/2}B^{1/2} + 4A^{1/3}B^{3/4} + B = (A^{1/4}+B^{1/4})^4\,. \quad \square$$

We can now prove Lemma 4.1. Let $\alpha_0 = \langle u, x_0\rangle \in \mathbb{R}[u]$. By construction, the polynomial identity $\|Pu\|_4^4 = \|\alpha_0 x_0 + P'u\|_4^4$ holds over $\mathbb{R}[u]$. By the triangle inequality for pseudodistribution ℓ_4 norm, for $A = \tilde{\mathbb{E}}\alpha_0^4\|x_0\|_4^4$ and $B = \tilde{\mathbb{E}}\|P'u\|_4^4$

$$\left(\tfrac{1}{\mu_0}\right)^{1/4} = \left(\tilde{\mathbb{E}}\|Pu\|_4^4\right)^{1/4} \le A^{1/4} + B^{1/4}$$

By the premises of the lemma, $A = \tilde{\mathbb{E}}\alpha_0^4/\mu_0$ and $B \le 1/\mu'$. Together with the previous bound, it follows that $(\tilde{\mathbb{E}}\alpha_0^4)^{1/4} \ge 1 - (\mu_0/\mu')^{1/4}$. Since $\alpha_0^2 \preccurlyeq \|u\|_2^2$ and $\{u\}$ satisfies $\|u\|_2^2 = 1$, we have $\tilde{\mathbb{E}}\alpha_0^2 \ge \tilde{\mathbb{E}}\alpha_0^4 \ge 1 - 4(\mu_0/\mu')^{1/4}$. Finally, using $\|u - x^{(0)}0\|_2^2 = \|u\|_2^2 - \alpha_0^2$, we derive the desired bound $\tilde{\mathbb{E}}\|u - x^{(0)}\|_2^2 = 1 - \tilde{\mathbb{E}}\alpha_0^2 \le 4(\mu_0/\mu')^{1/4}$ thus establishing Lemma 4.5 and Theorem 4.1. $\square$

4.2. Sparse dictionary learning. A *κ-overcomplete dictionary* is a matrix $A \in \mathbb{R}^{n\times m}$ with $\kappa = m/n \ge 1$ and isotropic unit vectors as columns (so that $\|A^\top u\|_2^2 = \kappa\|u\|_2^2$). We say a distribution $\{x\}$ over $\mathbb{R}^m$ is (d,τ)-nice if it satisfies $\mathbb{E}_i\, x_i^d = 1$ and $\mathbb{E}_i\, x_i^2 x_j^2 \le \tau$ for all $i \ne j \in [m]$, and it satisfies that non-square monomial degree-d moments vanish so that $\mathbb{E}\, x^\alpha = 0$ for all non-square degree-d monomials x^α, where $x^\alpha = \prod x_i^{\alpha_i}$ for $\alpha \in \mathbb{Z}^n$. For $\tau = o(1)$, a nice distribution satisfies that $\mathbb{E}\,\frac{1}{m}\sum_i x_i^4 \gg \left(\frac{1}{m}\sum_i x_i^2\right)^2$ which means that it is approximately sparse in the sense that the square of the entries of x has large variance (which means that few of the entries have very big magnitude compared to the rest).

Theorem 4.6. *For every $\varepsilon > 0$ and $\kappa \ge 1$, there exists d and τ and a quasipolynomial-time algorithm algorithm for SPARSE DICTIONARY LEARNING with the following guarantees: Suppose the input consists of $n^{O(1)}$ independent samples[17] from a distribution $\{y = Ax\}$ over $\mathbb{R}^n$, where $A \in \mathbb{R}^{n\times m}$ is a κ-overcomplete dictionary and the distribution $\{x\}$ over $\mathbb{R}^m$ is (d,τ)-nice. Then, with high probability, the algorithm outputs a set of vectors with Hausdorff distance[18] at most ε from the set of columns of A.*

Encoding as a system of polynomial equations. Let $y^{(}0), \dots, y^{(}R$ be independent samples from the distribution $\{y = Ax\}$. Then, we consider the polynomial $P = \frac{1}{R}\sum_i \langle y^{(}i, u\rangle^d$ in $\mathbb{R}[u]_d$. Using the properties of nice distributions, a direct computation shows that with high probability P satisfies the relation

$$\|A^\top u\|_d^d - \tau\|u\|_2^d \preccurlyeq P \preccurlyeq \|A^\top u\|_d^d + \tau\|u\|_2^d\,.$$

(Here, we are omitting some constant factors, depending on d, that are not important for the following discussion.) It follows that $P(a^{(}i) = 1 \pm \tau$ for every column $a^{(}i$ of A. It's also

[17] Here, we also make the mild assumption that the degree-$2d$ moments of x are bounded by $n^{O(1)}$.

[18] The *Hausdorff distance* between two sets of vectors upper bounds the maximum distance of a point in one of the sets to its closest point in the other set. Due to the innate symmetry of the sparse dictionary problem (replacing a column $a^{(}i$ of A by $-a^{(}i$ might not affect the input distribution), we measure the Hausdorff distance after symmetrizing the sets, i.e., replacing the set S by $S \cup -S$.

not hard to show that every unit vector a^* with $P(a^*) \approx 1$ is close to one of the columns of A. (Indeed, every unit vector satisfies $P(a^*) \leq \max_i \langle a^{\circ i}, a^* \rangle^{d-2} \kappa + \tau$. Therefore, $P(a^*) \approx 1$ implies that $\langle a^{\circ i}, a^* \rangle^2 \geq \kappa^{-\Omega(1/d)}$, which is close to 1 for $d \gg \log \kappa$.) What we will show is that pseudodistributions of degree $O(\log n)$ allow us to find all such vectors.

Why does the sum-of-squares method work? In the following, $\varepsilon > 0$ and $\kappa \geq 1$ are arbitrary constants that determine constants $d = d(\varepsilon, \kappa) \geq 1$ and $\tau = \tau(\varepsilon, \kappa) > 0$ (as in the theorem).

Lemma 4.7. *Let $P \in \mathbb{R}[u]$ be a degree-d polynomial with $\pm(P - \|A^\top u\|_d^d) \preccurlyeq \tau \|u\|_2^d$ for some κ-overcomplete dictionary A. Let $\mathcal{D}$ be a degree-$O(\log n)$ pseudodistribution that satisfies the constraints $\{\|u\|_2^2 = 1\}$ and $\{P(u) = 1 - \tau\}$. Let $W \in \mathbb{R}[u]$ be a product of $O(\log n)$ random linear forms*[19]*. Then, with probability at least $n^{-O(1)}$ over the choice of W, there exists a column $a^{\circ i}$ of A such that*

$$\tfrac{1}{\tilde{\mathbb{E}}_{\mathcal{D}} W^2} \tilde{\mathbb{E}}_{\mathcal{D}} W^2 \cdot \big(\|u\|^2 - \langle a^{\circ i}, u \rangle^2 \big) \leq \varepsilon \,.$$

If $\tilde{\mathbb{E}}_{\mathcal{D}}$ is a pseudoexpectation operator, then $\tilde{\mathbb{E}}_{\mathcal{D}'} \colon P \mapsto \tilde{\mathbb{E}} W^2 P / \tilde{\mathbb{E}} W^2$ is also a pseudoexpectation operator (as it satisfies linearity, normalization, and nonnegativity). (This transformation corresponds to reweighing the pseudodistribution $\mathcal{D}$ by the polynomial W^2.) Hence, the conclusion of the lemma gives us a new pseudodistribution $\mathcal{D}'$ such that $\tilde{\mathbb{E}}_{\mathcal{D}'} \|u\|_2^2 - \langle a^{\circ i}, u \rangle^2 \leq \varepsilon$. Therefore, if we sample a Gaussian vector a^* with the same quadratic moments as $\mathcal{D}'$, it satisfies $\|a^*\|_2^2 - \langle a^{\circ i}, a^* \rangle^2 \leq 4\varepsilon$ with probability at least $3/4$. At the same time, it satisfies $\|a^*\|^2 \geq 1/4$ with probability at least $1/2$. Taking these bounds together, a^* satisfies $\langle a^{\circ i}, a^* \rangle^2 \geq (1 - 16\varepsilon) \|a^*\|^2$ with probability at least $1/4$.

Lemma 4.7 allows us to reconstruct one of the columns of A. Using similar ideas, we can iterate this argument and recover one-by-one all columns of A. We omit the proof of Lemma 4.7, but the idea behind it is to first give an SOS proof version of our argument above that maximizers of P must be close to one of the $a^{(i)}$'s. We then note that if a distribution $\mathcal{D}$ is supported (up to noise) on at most m different vectors, then we can essentially isolate one of these vectors by re-weighing $\mathcal{D}$ using the product of the squares of $O(\log m)$ random linear forms.

It turns out, this latter argument has a low degree SOS proof as well, which means that in our case that given $\mathcal{D}$ satisfying the constraint $\{P(u) = 1 - \tau\}$, we can isolate one of the $a^{\circ i}$'s even when $\mathcal{D}$ is not an actual distribution but merely a pseudodistribution.

5. Hypercontractive norms and small-set expansion

So far we have discussed the Small-Set Expansion Hypothesis and the Sum of Squares algorithm. We now discuss how these two notions are related. One connection, mentioned before, is that the SSEH predicts that in many settings the guarantees of the degree-2 SOS algorithm are best possible, and so in particular it means that going from degree 2 to say degree 100 should not give any substantial improvement in terms of guarantees. Another, perhaps more meaningful connection is that there is a candidate approach for refuting the SSEH using the SOS algorithm. At the heart of this approach is the following observation:

[19] Here, a random linear form means a polynomial $\langle u, v \rangle \in \mathbb{R}[u]$ where v is a random unit vector in $\mathbb{R}^n$.

The small-set expansion problem is a special case of the problem of finding "sparse" vectors in a linear subspace.

This may seem strange, as a priori, the following two problem seem completely unrelated: **(i)** Given a graph $G = (V, E)$, find a "small" subset $S \subseteq V$ with low expansion $\phi_G(S)$, and **(ii)** Given a subspace $W \subseteq \mathbb{R}^n$, find a "sparse" vector in W. The former is a combinatorial problem on graphs, and the latter a geometric problem on subspaces. However, for the right notions of "small" and "sparse", these turn out to be essentially the same problem. Intuitively, the reason is the following: the expansion of a set S is proportional to the quantity $x^\top Lx$ where x is the characteristic vector of S (i.e. x_i equals 1 if $i \in S$ and equals 0 otherwise), and L is the *Laplacian matrix* of G (defined as $L = I - d^{-1}A$ where I is the identity, d is the degree, and A is G's adjacency matrix). Let $v_1, \ldots, v_n$ be the eigenvectors of L and $\lambda_1, \ldots, \lambda_n$ the corresponding eigenvalues. Then $x^\top Lx = \sum_{i=1}^n \lambda_i \langle v_i, x\rangle^2$.

Therefore if $x^\top Lx$ is smaller than $\varphi\|x\|^2$ and c is a large enough constant, then most of the mass of x is contained in the subspace $W = \mathrm{Span}\{v_i : \lambda_i \leq c\varphi\}$. Since S is small, x is sparse, and so we see that there is a sparse vector that is "almost" contained in W. Moreover, by projecting x into W we can also find a "sparse" vector that is actually contained in W, if we allow a slightly softer notion of "sparseness", that instead of stipulating that most coordinates are zero, only requires that the distribution of coordinates is very "spiky" in the sense that most of its mass is dominated by the few "heavy hitters".

Concretely, for $p > 1$ and $\delta \in (0, 1)$, we say that a vector $x \in \mathbb{R}^n$ is (δ, p)-*sparse* if $\mathbb{E}_i\, x_i^{2p} \geq \delta^{1-p}(\mathbb{E}_i\, x_i^2)^p$. Note that a characteristic vector of a set of measure δ is (δ, p)-sparse for any p. The relation between small-set-expansion and finding sparse vectors in a subspace is captured by the following theorem:

Theorem 5.1 (Hypercontractivity and small-set expansion [12], informal statement). *Let $G = (V, E)$ be a d-regular graph with Laplacian L. Then for every $p \geq 2$ and $\varphi \in (0, 1)$,*

(1) (Non-expanding small sets imply sparse vectors.) *If there exists $S \subseteq V$ with $|S| = o(|V|)$ and $\phi_G(S) \leq \varphi$ then there exists an $(o(1), p)$-sparse vector $x \in W_{\leq \varphi + o(1)}$ where for every λ, $W_{\leq\lambda}$ denotes the span of the eigenvectors of L with eigenvalue smaller than λ.*

(2) (Sparse vectors imply non-expanding small sets.) *If there exists a $(o(1), p)$-sparse vector $x \in W_{\leq\varphi}$, then there exists $S \subseteq V$ with $|S| = o(|V|)$ and $\phi_G(S) \leq \rho$ for some constant $\rho < 1$ depending on φ.*

The first direction of Theorem 5.1 follows from the above reasoning, and was known before the work of [12]. The second direction is harder, and we omit the proof here. The theorem reduces the question of determining whether there for small sets S, the minimum of $\phi_G(S)$ is close to one or close to zero, into the question of bounding the maximum of $\mathbb{E}_i\, x_i^{2p}$ over all unit vectors in some subspace. The latter question is a polynomial optimization problem of the type the SOS algorithm is designed for! Thus, we see that we could potentially resolve the SSEH if we could answer the following question:

What is the degree of SOS proofs needed to certify that the $2p$-norm is bounded for all (Euclidean norm) unit vectors in some subspace W?

We still don't know the answer to this question in full generality, but we do have some interesting special cases. Lemma 4.2 of Section 4.1 implies that if W is a random subspace of

dimension $\ll \sqrt{n}$ then we can certify that $\mathbb{E}_i x_i^4 \leq O(\mathbb{E}_i x_i^2)^2$ for all $x \in W$ via a degree-4 SOS proof. This is optimal, as the 4-norm simply won't be bounded for dimensions larger than $\sqrt{n}$:

Lemma 5.2. *Let $W \subseteq \mathbb{R}^n$ have dimension d and $p \geq 2$, then there exists a unit vector $x \in W$ such that*

$$\mathbb{E}_i x_i^{2p} \geq \tfrac{d^p}{n} (\mathbb{E}_i x_i^2)^p$$

Hence in particular any subspace of dimension $d \gg n^{1/p}$ contains a $(o(1), p)$-sparse vector.

Proof of Lemma 5.2. Let P be the matrix corresponding to the projection operator to the subspace W. Note that P has d eigenvalues equalling 1 and the rest equal 0, and hence $\mathrm{Tr}(P) = d$ and the Frobenius norm squared of P, defined as $\sum P_{i,j}^2$, also equals d. Let $x^i = Pe^i$ where e^i is the i^{th} standard basis vector. Then $\sum x_i^i$ is the trace of P which equals d and hence using Cauchy-Schwarz

$$\sum (x_i^i)^2 \geq \frac{1}{n} \left(\sum x_i \right)^2 = \frac{\mathrm{Tr}(P)^2}{n} = \frac{d^2}{n} \,.$$

On the other hand,

$$\sum_i \sum_j (x^i)_j^2 = \sum_{i,j} (Pe^i)_j^2 = \sum P_{i,j}^2 = d \,.$$

Therefore, by the inequality $(\sum a_i)/(\sum b_i) \leq \max a_i/b_i$, there exists an i such that if we let $x = x^i$ then $x_i^2 \geq \frac{d}{n} \sum_j x_j^2 = d\, \mathbb{E}\, x_j^2$. Hence, just the contribution of the i^{th} coordinate to the expectation achieves $\mathbb{E}_j x_j^{2p} \geq \frac{d^p}{n} \left(\mathbb{E}_j x_j^2 \right)^p$. □

Lemma 5.2 implies the following corollary:

Corollary 5.3. *Let $p, n \in \mathbb{N}$, and W be subspace of $\mathbb{R}^n$. If $\mathbb{E}_i x_i^{2p} \leq O(\mathbb{E}_i x_i^2)^p)$, then there is an $O(n^{1/p})$-degree SOS proof for this fact. (The constants in the $O(\cdot)$ notation can depend on p but not on n.)*

Proof sketch. By Lemma 5.2, the condition implies that $d = \dim W \leq O(n^{1/p})$, and it is known that approximately bounding a degree-$O(1)$ polynomial on the d-dimensional sphere requires an SOS proof of at most $O(d)$ degree (e.g., see [22] and the references therein). □

Combining Corollary 5.3 with Theorem 5.1 implies that for every ε, δ there exists some τ (tending to zero with ε), such that if we want to distinguish between the case that an n-vertex graph G satisfies $\phi_G(S) \leq \varepsilon$ for every $|S| \leq \delta n$, and the case that there exists some S of size at most δn with $\phi_G(S) \geq 1 - \varepsilon$, then we can do so using a degree n^τ SOS proofs, and hence in $\exp(O(n^\tau))$ time. This is much better than the trivial $\binom{n}{\delta n}$ time algorithm that enumerates all possible sets. Similar ideas can be used to achieve an algorithm with a similar running time for the problem underlying the Unique Games Conjecture [5]. If these algorithms could be improved so the exponent τ tends to zero with n for a fixed ε, this would essentially refute the SSEH and UGC.

Thus, the question is whether Corollary 5.3 is the best we could do. As we've seen, Lemma 4.2 shows that for random subspaces we can do much better, namely certify the bound with a constant degree proof. Two other results are known of that flavor. Barak, Kelner and Steurer [15] showed that if a d-dimensional subspace W does not contain a

$(\delta, 2)$-sparse vector, then there is an $O(1)$-degree SOS proof that it does not contain (or even almost contains) a vector with $O(\frac{\delta n}{d^{1/3}})$ nonzero coordinates. If the dependence on d could be eliminated (even at a significant cost to the degree), then this would also refute the SSEH. Barak, Brandão, Harrow, Kelner, Steurer and Zhou [12] gave an $O(1)$-degree SOS proof for the so-called "Bonami-Beckner-Gross $(2, 4)$ hypercontractivity theorem" (see [43, Chap. 9]). This is the statement that for every constant k, the subspace $W_k \subseteq \mathbb{R}^{2^t}$ containing the evaluations of all degree $\leq k$ polynomials on the points $\{\pm 1\}^t$ does not contain an $(o(1), 2)$-sparse vector, and specifically satisfies for all $x \in W_k$,

$$\mathbb{E}\, x_i^4 \leq 9^k (\mathbb{E}\, x_i^2)^2 \,. \tag{5.1}$$

On its own this might not seem so impressive, as this is just one particular subspace. However, this particular subspace underlies much of the evidence that has been offered so far in support of both the UGC and SSEH conjectures. The main evidence for the UGC/SSEH consists of several papers such as [13, 34, 35, 47] that verified the predictions of these conjectures by proving that various natural algorithms indeed fail to solve some of the computational problems that are hard if the conjectures are true. These results all have the form of coming up with a "hard instance" G on which some algorithm $\mathcal{A}$ fails, and so to prove such a result one needs to do two things: **(i)** compute (or bound) the true value of the parameter on G, and **(ii)** show that the value that $\mathcal{A}$ outputs on G is (sufficiently) different than this true value. It turns out that all of these papers, the proof of **(i)** can be formulated as low degree SOS proof, and in fact the heart of these proofs is the bound (5.1). Therefore, the results of [12] showed that all these "hard instances" can in fact be solved by the SOS algorithm using a constant degree. This means that at the moment, we don't even have any example of an instance for the problems underlying the SSEH and UGC that can be reasonably *conjectured* (let alone proved) hard for the constant degree SOS algorithm. This does not mean that such instances do not exist, but is suggestive that we have not yet seen the last algorithmic word on this question.

Acknowledgments. We thank Amir Ali Ahmadi for providing us with references on the diverse applications of the SOS method.

References

[1] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, *Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization*, preprint arXiv:1310.7991, 2013.

[2] N. Alon, *Decomposition of the completer-graph into completer-partiter-graphs*, Graphs and Combinatorics, **2** (1) (1986), 95–100.

[3] N. Alon and V. D. Milman, λ_1*, Isoperimetric inequalities for graphs, and superconcentrators*, J. Comb. Theory, Ser. B, **38**(1) (1985), 73–88.

[4] C. Ambühl, M. Mastrolilli, and O. Svensson, *Inapproximability Results for Maximum Edge Biclique, Minimum Linear Arrangement, and Sparsest Cut*, SIAM J. Comput., **40**(2) (2011), 567–596.

[5] S. Arora, B. Barak, and D. Steurer, Subexponential Algorithms for Unique Games and Related Problems. In FOCS, pp. 563–572, 2010.

[6] S. Arora, R. Ge, and A. Moitra, *New Algorithms for Learning Incoherent and Overcomplete Dictionaries*, preprint arXiv:1308.6723, 2013.

[7] S. Arora, S. Rao, and U. V. Vazirani, *Expander flows, geometric embeddings and graph partitioning*, J. ACM, **56**(2) (2009).

[8] E. Artin, *Über die zerlegung definiter funktionen in quadrate*, In Abhandlungen aus dem mathematischen Seminar der Universität Hamburg, volume 5, Springer, 1927, pp. 100–115.

[9] B. Barak, *Truth vs. Proof in Computational Complexity*, Bulletin of the European Association for Theoretical Computer Science, (108), October 2012.

[10] ———, *Fun and Games with Sums of Squares*, Feb 2014, Windows on Theory blog, http://windowsontheory.org

[11] ———, *Structure vs. Combinatorics in Computational Complexity*, Bulletin of the EATCS, (112):115–126, February 2014.

[12] B. Barak, F. G. S. L. Brandão, A. W. Harrow, J. A. Kelner, D. Steurer, and Y. Zhou, *Hypercontractivity, sum-of-squares proofs, and their applications* , In STOC, pp. 307–326, 2012.

[13] B. Barak, P. Gopalan, J. Håstad, R. Meka, P. Raghavendra, and D. Steurer, *Making the Long Code Shorter*, In FOCS, pp. 370–379, 2012.

[14] B. Barak, J. Kelner, and D. Steurer, *Dictionary Learning via the Sum-of-Squares Method*, Unpublished manuscript, 2014.

[15] ———, *Rounding Sum of Squares Relaxations*, In STOC, 2014.

[16] G. Blekherman, P. A. Parrilo, and R. R. Thomas, *Semidefinite optimization and convex algebraic geometry*, volume 13, Siam, 2013.

[17] J. Cheeger, *A lower bound for the smallest eigenvalue of the Laplacian*, Problems in analysis, **625** (1970), 195–199.

[18] J. Chen, X. Huang, I. A. Kanj, and G. Xia, *Strong computational lower bounds via parameterized complexity*, Journal of Computer and System Sciences, **72**(8) 2006, 1346–1367.

[19] P. Comon, *Independent component analysis, a new concept?* Signal processing, **36**(3) (1994), 287–314.

[20] L. Demanet and P. Hand, *Recovering the Sparsest Element in a Subspace*, preprint arXiv:1310.1654, Oct 2013.

[21] J. Dodziuk, *Difference equations, isoperimetric inequality and transience of certain random walks*, Transactions of the American Mathematical Society, **284**(2) (1984), 787–794.

[22] A. C. Doherty and S. Wehner, *Convergence of SDP hierarchies for polynomial optimization on the hypersphere*, preprint arXiv:1210.5048, 2012.

[23] R. G. Downey and M. R. Fellows, *Fixed-parameter tractability and completeness II: On completeness $W[1]$*, Theoretical Computer Science, **141**(1) (1995), 109–131.

[24] M. X. Goemans and D. P. Williamson, *Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming*, J. ACM, **42**(6) (1995), 1115–1145.

[25] N. Goyal, S. Vempala, and Y. Xiao, *Fourier PCA*, In STOC, 2014, Also available as arXiv report arXiv:1306.5825.

[26] D. Grigoriev, *Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity*, Theor. Comput. Sci., **259**(1-2) (2001), 613–622.

[27] D. Grigoriev and N. Vorobjov, *Complexity of Null-and Positivstellensatz proofs*, Annals of Pure and Applied Logic, **113**(1) (2001), 153–160.

[28] A. Grothendieck, *Résumé de la théorie métrique des produits tensoriels topologiques* , Bol. Soc. Mat. Sao Paulo, **8**(1-79) (1953), 88.

[29] J. Håstad, *Clique is Hard to Approximate Within $n^{1-\epsilon}$*, In FOCS, pp. 627–636, 1996.

[30] S. Khot, *Improved Inaproximability Results for MaxClique, Chromatic Number and Approximate Graph Coloring*, In FOCS, pp. 600–609, 2001.

[31] ———, *On the Power of Unique 2-Prover 1-Round Games*, In IEEE Conference on Computational Complexity, p. 25, 2002.

[32] ———, *Inapproximability of np-complete problems, discrete fourier analysis, and geometry*, In International Congress of Mathematics, volume 5, 2010.

[33] ———, *On the Unique Games Conjecture (Invited Survey)*, In 2012 IEEE 27th Conference on Computational Complexity, pp. 99–121, 2010.

[34] S. Khot and R. Saket, *SDP Integrality Gaps with Local ℓ_1-Embeddability*, In FOCS, pp. 565–574, 2009.

[35] S. Khot and N. K. Vishnoi, *The Unique Games Conjecture, Integrality Gap for Cut Problems and Embeddability of Negative Type Metrics into ℓ_1*, In FOCS, pp. 53–62, 2005.

[36] J.-L. Krivine, *Anneaux préordonnés* Journal d'analyse mathématique, **12**(1) (1964), 307–326.

[37] J. B. Lasserre, *Global Optimization with Polynomials and the Problem of Moments*, SIAM Journal on Optimization, **11**(3) (2001), 796–817.

[38] M. Laurent, *A Comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre Relaxations for 0-1 Programming*, Math. Oper. Res., **28**(3) (2003), 470–496.

[39] ———, *Sums of squares, moment matrices and optimization over polynomials*, In Emerging applications of algebraic geometry, Springer, 2009, pp. 157–270.

[40] L. Lovász, *On the Shannon capacity of a graph*, Information Theory, IEEE Transactions on, **25**(1) (1979), 1–7.

[41] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and 0-1 optimization*, SIAM Journal on Optimization, **1**(2) (1991), 166–190.

[42] Y. Nesterov, *Squared functional systems and optimization problems*, High performance optimization, **13** (2000), 405–440.

[43] R. O'Donnell, *Analysis of Boolean Functions*, Cambridge University Press, May 2014.

[44] B. A. Olshausen and D. J. Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, Nature, **381**(6583) (1996), 607–609.

[45] P. A. Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, PhD thesis, California Institute of Technology, 2000.

[46] P. Raghavendra, *Optimal algorithms and inapproximability results for every CSP?*, In STOC, pp. 245–254, 2008.

[47] P. Raghavendra and D. Steurer, *Integrality Gaps for Strong SDP Relaxations of UNIQUE GAMES*, In FOCS, pp. 575–585, 2009.

[48] _____, *Towards computing the Grothendieck constant* In SODA, pp. 525–534, 2009.

[49] _____, *Graph expansion and the unique games conjecture* In STOC, pp. 755–764, 2010.

[50] P. Raghavendra, D. Steurer, and P. Tetali, *Approximations for the isoperimetric and spectral profile of graphs and related parameters*, In STOC, pp. 631–640, 2010.

[51] P. Raghavendra, D. Steurer, and M. Tulsiani, *Reductions between Expansion Problems* In IEEE Conference on Computational Complexity, pp. 64–73, 2012.

[52] B. Reznick, *Some concrete aspects of Hilbert's 17th problem*, Contemporary Mathematics, **253** (2000), 251–272.

[53] G. Schoenebeck, *Linear Level Lasserre Lower Bounds for Certain k-CSPs*, In FOCS, pp. 593–602, 2008.

[54] H. D. Sherali and W. P. Adams, *A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems* SIAM Journal on Discrete Mathematics, **3**(3) (1990), 411–430.

[55] N. Shor, *An approach to obtaining global extremums in polynomial mathematical programming problems*, Cybernetics and Systems Analysis, **23**(5) (1987), 695–700.

[56] D. A. Spielman, H. Wang, and J. Wright, *Exact Recovery of Sparsely-Used Dictionaries*, Journal of Machine Learning Research - Proceedings Track, 23:37.1–37.18, 2012.

[57] G. Stengle, *A Nullstellensatz and a Positivstellensatz in semialgebraic geometry*, Mathematische Annalen, **207**(2) (1974), 87–97.

[58] D. Steurer, *Fast SDP Algorithms for Constraint Satisfaction Problems*, In SODA, pp. 684–697, 2010.

[59] L. Trevisan, *On Khot's Unique Games Conjecture*, Bulletin (New Series) of the American Mathematical Society, **49**(1) (2012).

[60] M. Tulsiani, *CSP gaps and reductions in the lasserre hierarchy*, In STOC, pp. 303–312, 2009.

Microsoft, 1 Memorial Drive, Cambridge, MA 02142
E-mail: info@boazbarak.org

Department of Computer Science, Cornell University, Ithaca NY 14853
E-mail: dsteurer@cs.cornell.edu

Interactive information and coding theory

Mark Braverman

Abstract. We give a high-level overview of recent developments in interactive information and coding theory. These include developments involving interactive noiseless coding and interactive error-correction. The overview is primarily focused on developments related to complexity-theoretic applications, although the broader context and agenda are also set out. As the present paper is an extended abstract, the vast majority of proofs and technical details are omitted, and can be found in the respective publications and preprints.

Mathematics Subject Classification (2010). Primary 94A15; Secondary 68Q99.

Keywords. Coding theory, communication complexity, information complexity, interactive computation.

1. Introduction

1.1. A high-level overview of information and coding theory. We begin with a very high-level overview of information and coding theory. This is an enormous field of study, with subareas dealing with questions ranging from foundations of probability and statistics to applied wireless transmission systems. We will focus only on some of the very basic foundational aspects, which were set forth by Shannon in the late 1940s, or shortly after. The goal will be to try and translate those to interactive communication settings, of the type that is used in theoretical computer science. This program is only very partially complete, but some of the early results are promising. While our overview of information and coding theory in this section focuses on fairly simple facts, we present those in some detail nonetheless, as they will be used as a scaffold for the interactive coding discussion. A thorough introduction into modern information theory is given in [15].

Noiseless coding. Classical information theory studies the setting where one terminal (Alice) wants to transmit information over a channel to another terminal (Bob). Two of the most important original contributions by Shannon are the *Noiseless Coding* (or Source Coding) Theorem and the *Noisy Coding* (or Channel Coding) Theorem. The Source Coding Theorem asserts that the cost of Alice transmitting n i.i.d. copies of a discrete random variable X to Bob *over a noiseless channel* scales as Shannon's entropy $H(X)$ as $n \to \infty$[1]:

$$H(X) = \sum_{x \in \mathrm{supp}(X)} \Pr[X = x] \log \frac{1}{\Pr[X = x]}. \tag{1.1}$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

If we denote by X^n the concatenation of n independent samples from X, and by $C(Y)$ the (expected) number of bits needed for Alice to transmit a sample of random variable Y to Bob, then the Source Coding Theorem asserts that[2]

$$\lim_{n\to\infty} \frac{C(X^n)}{n} = H(X). \tag{1.2}$$

This fact can be viewed as the operational definition of entropy, i.e. one that is grounded in reality. Whereas definition (1.1) may appear artificial, (1.2) implies that it is the right one, since it connects to the "natural" quantity $C(X^n)$. Another indirect piece of evidence indicating that $H(X)$ is a natural quantity is its additivity property:

$$H(X^n) = n \cdot H(X), \tag{1.3}$$

and more generally, if XY is the concatenation of random variables X and Y, then $H(XY) = H(X) + H(Y)$ whenever X and Y are independent. Note that it is not hard to see that (1.3) fails to hold for $C(X)$, making $H(X)$ a "nicer" quantity to deal with than $C(X)$. Huffman coding (1.11) below blurs the distinction between the two, as they only differ by at most one additive bit, but we will return to it later in the analogous distinction between communication complexity and information complexity.

Noisy coding. So far we assumed a noiseless channel — bits sent over the channel by Alice are received by Bob unaltered. If the channel is *noisy*, that is, messages sent over the channel may get corrupted, then clearly some redundancy in transmission is necessary. Abstractly, the task of *coding* is the task of converting the message being sent into symbols to be transmitted over the channel, in a way that allows the original message to be recovered from what has been transmitted by the channel. The important considerations for how good a code is are the type (and amount) of errors it can withstand and still accomplish the transmission successfully, and the rate by which the error-correcting encoding enlarges the message being transmitted.

Shannon's Noisy-Channel Coding Theorem was first to address the noisy coding scenario theoretically. The most important insight from that theorem is that, at least in the limit, the ability of a channel to conduct information — defined formally as Shannon's channel capacity — can be decoupled from the content being transmitted over the channel. Informally, for a memoryless channel $\mathcal{C}$ one can define its capacity $\mathsf{cap}(\mathcal{C})$ as "how many bits of information is one utilization of $\mathcal{C}$ (i.e. one transmission over $\mathcal{C}$) worth?". For any X, if we denote by $C_{\mathcal{C}}(X^n)$ the expected number of utilizations of channel $\mathcal{C}$ needed to transmit n independent samples of X (except with negligible error), then

$$\lim_{n\to\infty} \frac{C_{\mathcal{C}}(X^n)}{n} = \frac{H(X)}{\mathsf{cap}(\mathcal{C})}. \tag{1.4}$$

This means, conveniently, that one can study properties of channels separately from properties of what is being transmitted over the channels. The information-theoretic quantities needed to express $\mathsf{cap}(\mathcal{C})$ are *conditional entropy* and *mutual information*. While these are

[1]All logs in this paper are base-2.

[2]In fact, Shannon's Source Coding Theorem asserts that due to concentration the *worst case* communication cost scales as $H(X)$ as well, if we allow negligible error. We ignore this stronger statement at the present level of abstraction.

standard basic notions in information theory, we will define them here, to keep the exposition accessible. For a pair of random variables X and Y, the conditional entropy $H(X|Y)$ can be thought of as the amount of uncertainty remaining in X for someone who knows Y:

$$H(X|Y) := H(XY) - H(Y) = \mathbf{E}_{y\sim Y} H(X|Y = y). \tag{1.5}$$

In the extreme case where X and Y are independent, we have $H(X|Y) = H(X)$. In the other extreme, when $X = Y$, we have $H(X|X) = 0$. The *mutual information* $I(X;Y)$ between two variables X and Y measures the amount of information revealing Y reveals about X, i.e. the reduction in X's entropy as a result of conditioning on Y. Thus

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(XY) = I(Y;X). \tag{1.6}$$

Conditional mutual information is defined similarly to conditional entropy:

$$I(X;Y|Z) := H(X|Z) - H(X|YZ) = I(Y;X|Z). \tag{1.7}$$

A very important property of conditional mutual information is the *chain rule*:

$$I(XY;Z|W) = I(X;Z|W) + I(Y;Z|WX) = I(Y;Z|W) + I(X;Z|WY). \tag{1.8}$$

An informal interpretation of (1.8) is that XY reveal about Z what X reveals about Z, plus what Y reveals about Z to someone who already knows X.

Abstractly, a memoryless channel (i.e. one where each utilization of the channel is independent of other utilization) can be viewed as a set of pairs of variables $(X, \mathcal{C}(X))$ where X is the signal the sender inputs to the channel, and $\mathcal{C}(X)$ is the output of the channel received on input X from the sender. If the channel is noiseless then $\mathcal{C}(X) = X$. Under this notation, the channel capacity of $\mathcal{C}$ is equal to

$$\mathsf{cap}(\mathcal{C}) = \sup_{Y} I(Y;\mathcal{C}(Y)). \tag{1.9}$$

In other words, it is the supremum over all input distributions of the amount of information preserved by the channel. The scenario just discussed is obviously a very simple one, but even in more elaborate settings issues surrounding coding transmissions over a noisy channel (at least when the noise is random) are very well understood. For example, for the *binary symmetric channel* BSC_ε that accepts bits $b \in \{0,1\}$ and outputs b with probability $1-\varepsilon$ and $\bar{b}$ with probability ε, the capacity is

$$\mathsf{cap}(BSC_\varepsilon) = 1 - H(\varepsilon) := 1 - (\varepsilon \log 1/\varepsilon + (1-\varepsilon)\log 1/(1-\varepsilon)). \tag{1.10}$$

One caveat is that mathematically striking characterizations such as above only become possible *in the limit*, where the size of the message we are trying to transmit over the channel — i.e. the block-length — grows to infinity. What happens for fixed block lengths, which we discuss next, is of course important for both practical and theoretical reasons, and it will be even more so in the interactive regime.

For noiseless coding in the one-way regime, it turns out that while $H(X)$ does not exactly equal the expected number of bits $C(X)$ needed to transmit a *single* sample from X, it is very close to it. For example, the classical Huffman's coding [25] implies that

$$H(X) \le C(X) < H(X) + 1, \tag{1.11}$$

where the "hard" direction of (1.11) is the upper bound. The upper bound showing that $C(X) < H(X) + 1$ is a *compression result*, showing how encode a message with low average information content (i.e. entropy) into a message with a low communication cost (i.e. number of bits in the transmission). Note that this result is much less "clean" than the limit result (1.2): in the amortized case the equality is exact, while in the one-shot case a gap is created. This gap is inevitable, if only for integrality reasons, but as we will see later, it becomes crucial in the interactive case.

Adversarial noise and list-decoding. So far we only discussed channels affected by randomized errors. A variant of the noisy regime where the situation appears mathematically much less clear is one where the errors on the channel are introduced *adversarially*. For example, an adversarial ε-error rate binary channel receives a string $S \in \{0,1\}^n$ of n bits, and outputs a string S' such that the Hamming distance $d_H(S,S') < \varepsilon n$, i.e. S' differs from S in at most an ε-fraction of positions. A coding scheme for this setting is a pair of encoding and decoding functions $E : \{0,1\}^m \to \{0,1\}^n$ and $D : \{0,1\}^n \to \{0,1\}^m$, respectively, such that for each $X \in \{0,1\}^m$ and each S' with $d_H(E(X),S') < \varepsilon n$, X is recovered correctly from S', i.e. $D(S') = X$. Clearly we want m to be as large as possible as a function of n. It turns out that such an encoding scheme is possible with $m = \Omega_\varepsilon(n)$ for each $\varepsilon < 1/4$ (and $\varepsilon < 1/2$ if the binary alphabet is replaced with an alphabet Σ of size[3] $|\Sigma| = O_\varepsilon(1)$). Unlike the random-noise case the exact optimal *rate* of the code, i.e. the largest achievable ratio of $\frac{m}{n}$ is unknown for the adversarial model. Clearly, the limit cannot exceed $\mathsf{cap}(BSC_\varepsilon)$, but it is bound to be lower, since correcting adversarial errors is much harder than randomized ones. *A priori* it is not even obvious that the adversarial channel capacity is a positive constant when $\varepsilon < 1/4$. Despite much work in the field [38, 45], even the basic binary channel capacity problem remains open, with a notorious gap between the Gilbert-Varshanov lower bound, and the Linear Programming upper bound [44].

A clear limitation of any error-correcting code, even over a large constant-size alphabet Σ, is that no decoding is possible when $\varepsilon \geq 1/2$: for two valid codewords X_1, X_2 and any encoding function E, there is a string S' such that $d_H(E(X_1),S') \leq n/2$ and $d_H(E(X_2),S') \leq n/2$, making decoding S' an impossible task (note that over a large constant-size alphabet, with a high probability one can recover from random errors of rate exceeding $1/2$). It turns out, however, that for any $\varepsilon < 1$, for $|\Sigma| = O_\varepsilon(1)$, it is possible to come up with a constant-rate *list-decoding* scheme: one where the decoding function $D(S')$ outputs a list of size $s = O_\varepsilon(1)$ of possible $X_1, \ldots, X_s$ such that these are the only possible X's satisfying $d_H(E(X),S') < (1-\varepsilon)n$. List decodable codes, first introduced in the 1950s [16, 47] have played an important role in a number of areas of theoretical computer science, a partial survey of which can be found in [23, 24].

1.2. Interactive computation models in complexity theory. In theoretical computer science interactive communication models are studied within the area of *communication complexity*. While communication complexity can be viewed as a direct extension of the study of non-interactive communication models which were discussed in the previous section, its development has been largely disjoint from the development of information theory, and the areas have not reconnected until fairly recently. This may be partially explained by the combinatorial nature of the tools most prevalent in theoretical computer science.

[3] The $O_\varepsilon(1)$ notation means a function that is bounded by a constant for each fixed ε.

Communication complexity was introduced by Yao in [48], and is the subject of the text [30]. It has found numerous applications for unconditional lower bounds in a variety of models of computation, including Turing machines, streaming, sketching, data structure lower bounds, and VLSI layout, to name a few. In the basic (two-party) setup, the two parties Alice and Bob are given inputs $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, respectively, and are required to compute a function $F(X,Y)$ of these inputs (i.e. both parties should know the answer in the end of the communication), while communicating over a noiseless binary channel. The parties are computationally unbounded, and their only goal is to minimize the number of bits transmitted in the process of computing $F(X,Y)$. In a typical setup F is a function $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$. Examples of functions commonly discussed and used include the Equality function $EQ_n(X,Y) := \mathbf{1}_{X=Y}(X,Y)$, and the Disjointness function

$$Disj_n(X,Y) := \bigwedge_{i=1}^{n} (\neg X_i \vee \neg Y_i). \tag{1.12}$$

We will return to these functions later in our discussion.

Of course, the (non-interactive) transmission problem can be viewed as just a special case of computing the function $P_x : \mathcal{X} \times \{\perp\} \to \mathcal{X}$, which maps $(X, \perp)$ to X. However, there are two important distinctions between the "flavors" of typical information theory results and communication complexity. Firstly, information theory is often concerned with coding results where block length — i.e. the number of copies of the communication task to be performed — goes to infinity. Recall, for example, that Shannon's Source Coding Theorem (1.2) gave Shannon's entropy as a closed-form expression for the amortized transmission cost of sending a growing number of samples X (this is often but not always the case, for example, the Huffman coding (1.11) result is not of this type). On the other hand, communication complexity more commonly studies the communication cost of computing a single copy of F. Secondly, as in the examples above, communication complexity often studies functions whose output is only a single bit or a small number of bits, thus "counting style" direct lower bound proofs rarely apply. Tools that have been successfully applied in communication complexity over the years include combinatorics, linear algebra, discrepancy theory, and only later classical information theory.

To make our discussion of communication complexity more technical, we will focus on the two-party setting. We briefly discuss the multi-party setting, which also has many important applications, but is generally much less well-understood, in the last section of the paper. The basic notion in communication complexity is that of a *communication protocol*. A communication protocol over a binary channel formalizes a conversation, where each message only depends on the input to the speaker and the conversation so far:

Definition 1.1. A (deterministic) protocol π for $F : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ is defined as a finite rooted binary tree, whose nodes correspond to partial communication transcripts, such that the two edges coming out of each vertex are labeled with a 0 and 1. Each leaf ℓ is labeled by an output value $f_\ell \in \{0,1\}$. Each internal node v is labeled by a player's name and either by a function $a_v : \mathcal{X} \to \{0,1\}$ or $b_v : \mathcal{Y} \to \{0,1\}$ corresponding to the next message of Alice or Bob, respectively.

The protocol $\pi(X,Y)$ is executed on a pair of inputs (X,Y) by starting from the root of the tree. At each internal node labeled by a_v the protocol follows the child $a_v(X)$ (corresponding to Alice sending a message), and similarly at each internal node labeled by b_v the protocol follows $b_v(Y)$. When a leaf ℓ is reached the protocol outputs f_ℓ.

By a slight abuse of notation, $\pi(X,Y)$ will denote both the transcript and the output of the protocol; which is the case will be clear from the context. The communication cost of a protocol is the depth of the corresponding protocol tree. A protocol *succeeds* on input (X,Y) if $\pi(X,Y) = F(X,Y)$. Its communication cost on this pair of inputs is the depth of the leaf reached by the execution. The *communication complexity* $CC(F)$ of a function F is the lowest attainable communication cost of a protocol that successfully computes F. In the case of deterministic communication we require the protocol to succeed on all inputs.

A deterministic communication protocol π induces a partition of the input space $\mathcal{X} \times \mathcal{Y}$ into sets S_ℓ by the leaf ℓ that $\pi(X,Y)$ reaches. Since at each step the next move of the protocol depends only on either X or Y alone, each S_ℓ is a combinatorial rectangle of the form $S_\ell = S_\ell^{\mathcal{X}} \times S_\ell^{\mathcal{Y}}$. This key combinatorial property is at the heart of many combinatorial communication complexity lower bounds. To give an example of such a simple combinatorial proof, consider the *rank* bound. Let $N = |\mathcal{X}|$, $M = |\mathcal{Y}|$, and consider the $N \times M$ matrix M_F over $\mathbb{R}$ whose (X,Y)-th entry is $F(X,Y)$. Each protocol π with leaf set $\mathcal{L}$ of size L, induces a partition of $\mathcal{X} \times \mathcal{Y}$ into combinatorial rectangles $\{S_\ell\}_{\ell \in \mathcal{L}}$. Let M_ℓ be the matrix whose entries are equal to $M_{X,Y}$ for $(X,Y) \in S_\ell$ and are 0 elsewhere. Since $\{S_\ell\}_{\ell \in \mathcal{L}}$ is a partition of $\mathcal{X} \times \mathcal{Y}$, we have $M_F = \sum_{\ell \in \mathcal{L}} M_\ell$. Assuming π is always correct, each M_ℓ is *monochromatic*, i.e. either all-0, or all-1 on S_ℓ, depending on the value of f_ℓ. Thus, $\text{rank}(M_\ell) \leq 1$, and

$$\text{rank}(M_F) \leq \sum_{\ell \in \mathcal{L}} \text{rank}(M_\ell) \leq L. \tag{1.13}$$

In fact, a stronger bound of $L - 1$ holds unless M_F is the trivial all-1 matrix. Thus any protocol computing F must have communication cost of at least $\log(\text{rank}(M_F) + 1)$, and it follows that the communication complexity of F is at least $\log(\text{rank}(M_F) + 1)$. As an example of an application, if $F = EQ_n$ is the Equality function, then $M_{EQ_n} = I_{2^n}$ is the identity matrix, and thus $CC(EQ_n) \geq n + 1$. In other words, the trivial protocol where Alice sends Bob her input X (n bits), and Bob responds whether $X = Y$ (1 bit), is optimal.

As in many other areas of theoretical computer science, there is much to be gained from randomization. For example, in practice, the Equality function does not require linear communication as Alice and Bob can just hash their inputs and compare the hash keys. The shorter protocol may return a false positive, but it is correct with high probability, and reduces the communication complexity from $n + 1$ to $O(\log n)$.

More generally, a randomized protocol is a protocol that tosses coins (i.e. accesses random bits), and produces the correct answer with high probability. The *distributional setting*, where there is a prior probability distribution μ on the inputs and the players need to output the correct answer with high probability with respect to μ is closely related to the randomized setting, as will be seen below. In the randomized setting there are two possible types of random coins. *Public coins* are generated at random and are accessible to both Alice and Bob at no communication cost. *Private coins* are coins generated privately by Alice and Bob, and are only accessible by the player who generated them. If Alice wants to share her coins with Bob, she needs to use the communication channel. In the context of communication complexity the pubic-coin model is clearly more powerful than the private coin one. Fortunately, the gap between the two is not very large [35], and can be mostly ignored. For convenience reasons, we will focus on the public-coin model.

The definition of a randomized public-coin communication protocol π_R is identical to Definition 1.1, except a public random string R is chosen at the beginning of the execution

of the randomized π_R, and all functions at the nodes of π_R may depend on R in addition to the respective input X or Y. We still require the answer f_ℓ to be unequivocally determined by the leaf ℓ alone. The communication cost $|\pi_R|$ of π_R is still its worst-case communication cost (for historic reasons; an average-case notion would also have been meaningful to discuss here).

The randomized communication complexity of F with error $\varepsilon > 0$ is given by

$$R_\varepsilon(F) := \min_{\pi_R : \forall X,Y \, \Pr_R[\pi_R(X,Y)=F(X,Y)] \geq 1-\varepsilon} |\pi_R|. \tag{1.14}$$

For a distribution μ on $\mathcal{X} \times \mathcal{Y}$ the *distributional* communication complexity $D_{\mu,\varepsilon}(F)$ is defined as the cost of the best protocol that achieves expected error ε with respect to μ. Note that in this case fixing public randomness R to a uniformly random value does not change (on average) the expected success probability of π_R with respect to μ. Therefore, without loss of generality, we may require π to be deterministic:

$$D_{\mu,\varepsilon}(F) := \min_{\pi : \mu\{X,Y:\ \pi(X,Y)=F(X,Y)\} \geq 1-\varepsilon} |\pi|. \tag{1.15}$$

It is easy to see that for all μ, $D_{\mu,\varepsilon}(F) \leq R_\varepsilon(F)$. By an elegant minimax argument [49], a partial converse is also true: for each F and ε, there is a distribution against which the distributional communication complexity is as high as the randomized:

$$R_\varepsilon(F) = \max_\mu D_{\mu,\varepsilon}(F). \tag{1.16}$$

For this reason, we will be able to discuss distributional and randomized communication complexity interchangeably.

How can one prove lower bounds for the randomized setting? This setting is much less restrictive than the deterministic one, making lower bounds more challenging. Given a function F, one can guess the hard distribution μ, and then try to lower bound the distributional communication complexity $D_{\mu,\varepsilon}(F)$ — that is, show that there is no low-communication protocol π that computes F with error $\leq \varepsilon$ with respect to μ. Such a protocol π of cost $k = |\pi|$ still induces a partition $\{S_\ell\}_{\ell \in \mathcal{L}}$ of the inputs according to the leaf they reach, with $L \leq 2^k$ and each S_ℓ a combinatorial rectangle. However, it is no longer the case that when we consider the corresponding submatrix M_ℓ of M_F it must be monochromatic — the output of π is allowed to be wrong on a fraction of S_ℓ, and thus for some inputs the output of π on S_ℓ may disagree with the value of F. Still, it should be true that for *most* leaves the value of F on S_ℓ is strongly biased one way or the other, since the contribution of S_ℓ to the error is

$$e(S_\ell) = \min\left(\mu(S_\ell \cap F^{-1}(0)), \mu(S_\ell \cap F^{-1}(1))\right). \tag{1.17}$$

In particular, a fruitful lower bound strategy is to show that all "large" rectangles with respect to μ have $e(S_\ell)/\mu(S_\ell) \gg \varepsilon$, and thus there must be many smaller rectangles — giving a lower bound on $L \leq 2^{|\pi|}$. One simple instantiation of this strategy is the *discrepancy* bound: for a distribution μ, the discrepancy $Disc_\mu(F)$ of F with respect to μ is the maximum over all combinatorial rectangles R of

$$Disc_\mu(R,F) := |\mu(F^{-1}(0) \cap R) - \mu(F^{-1}(1) \cap R)|.$$

In other words, if F has low discrepancy with respect to μ then only very small rectangles (as measured by μ) can be unbalanced. With some calculations, it can be shown that for all

$\varepsilon > 0$ (see [30] and references therein),

$$D_{\mu,\frac{1}{2}-\varepsilon}(F) \geq \log_2(2\varepsilon/Disc_\mu(F)). \tag{1.18}$$

Note that (1.18) not only says that if the discrepancy is low then the communication complexity is high, but also that it remains high even if we are only trying to gain a tiny advantage over random guessing in computing F! An example of a natural function to which the discrepancy method can be applied is the n-bit Inner Product function $IP_n(X,Y) = \langle X,Y\rangle \mod 2$. This simple discrepancy method can be generalized to a richer family of *corruption bounds* that can be viewed as combinatorial generalizations of the discrepancy bound. More on this method can be found in the survey [31].

One of the early successes of applying combinatorial methods in communication complexity was the proof that the *randomized* communication complexity of the set disjointness problem (1.12) is linear, $R_{1/4}(Disj_n) = \Theta(n)$. The first proof of this fact was given in the 1980s [26], and a much simpler proof was discovered soon after [41]. The proofs exhibit a specific distribution μ of inputs on which the distributional communication complexity $D_{\mu,1/4}(Disj_n)$ is $\Omega(n)$. Note that the uniform distribution would not be a great fit, since uniformly drawn sets are non-disjoint with a very high probability. It turns out that the following family of distributions μ is hard: select each coordinate pair (X_i,Y_i) i.i.d. from a distribution on $\{(0,0),(0,1),(1,0)\}$ (e.g. uniformly). This generates a distribution on pairs of disjoint sets. Now, with probability $1/2$ choose a uniformly random coordinate $i \in_U [n]$ and set $(X_i,Y_i) = (1,1)$. Thus, under μ, X and Y are disjoint with probability $1/2$.

Treating communication complexity as a generalization of one-way communication and applying information-theoretic machinery to it is a very natural approach (perhaps the most natural, given the success of information theory in communication theory). Interestingly, however, this is not how the field has evolved. In fact, the fairly recent survey [31] was able to present the vast majority of communication complexity results to its date without dealing with information theory at all. It is hard to speculate why this might be the case. One possible explanation is that the mathematical machinery needed to tackle the (much more complicated) interactive case from the information-theoretic angle wasn't available until the 1990s; another possible explanation is that linear algebra, linear programming duality, and combinatorics (the main tools in communication complexity lower bounds) are traditionally more central to theoretical computer science research and education than information theory.

A substantial amount of literature exists on communication complexity within the information theory community, see for example [36, 37] and references therein. The flavor of the results is usually different from the ones discussed above. In particular, there is much more focus on bounded-round communication, and significantly less focus on techniques for obtaining specific lower bounds for the communication complexity of specific functions such as the disjointness function. The most relevant work to our current discussion is a relatively recent line of work by Ishwar and Ma, which studied interactive amortized communication and obtained characterizations closely related to the ones discussed below [32, 33].

Within the theoretical computer science literature, in the context of communication complexity, information theoretic tools were explicitly introduced in [13] in the early 2000s for the simultaneous message model (i.e. 2 non-interactive rounds of communication). Building on this work, [1] developed tools for applying information theoretic reasoning to fully interactive communication, in particular giving an alternative (arguably, more intuitive) proof for the $\Omega(n)$ lower bound on the communication complexity of $Disj_n$. The motivating questions for [13], as well as for subsequent works developing information complexity, were the

direct sum [17] and *direct product* questions for (randomized) communication complexity.

In general, a direct sum theorem quantifies the cost of solving a problem F^n consisting of n sub-problems in terms of n and the cost of each sub-problem F. The value of such results to lower bounds is clear: a direct sum theorem, together with a lower bound on the (easier-to-reason-about) sub-problem, yields a lower bound on the composite problem (a process also known as hardness amplification). For example, the Karchmer-Wigderson program for boolean formula lower bounds can be completed via a (currently open) direct sum result for a certain communication model [27]. Direct product results further sharpen direct sum theorems by showing a "threshold phenomenon", where solving F^n with insufficient resources is shown to be impossible to achieve except with an exponentially small success probability. Classic results in complexity theory, such as Raz's Parallel Repetition Theorem [39] can be viewed as a direct product result.

In the next section, we will formally introduce information complexity, first as a generalization of Shannon's entropy to interactive tasks. We will then discuss its connections to the direct sum and product questions for randomized communication complexity, and to recent progress towards resolving these questions.

2. Noiseless coding and information complexity

Interactive information complexity. In this section we will work towards developing information complexity as the analogue of Shannon's entropy for interactive computation. It will sometimes be convenient to work with general *interactive two-party tasks* rather than just functions. A task $T(X, Y)$ is any action on inputs (X, Y) that can be performed by a protocol. $T(X, Y)$ can be though of as a set of distributions of outputs that are acceptable given an input (X, Y). Thus "computing $F(X, Y)$ correctly with probability $1 - \varepsilon$" is an example of a task, but there are examples of tasks that do not involve function or relation computation, for example "Alice and Bob need to sample strings A and B, respectively, distributed according to $(A, B) \sim \mu_{(X,Y)}$". For the purposes of the discussion, it suffices to think about T as the task of computing a function with some success probability. The communication complexity of a task T is then defined analogously to the communication complexity of functions. It is the least amount of communication needed to successfully perform the task $T(X, Y)$ by a communication protocol $\pi(X, Y)$.

The *information complexity* of a task T is defined as the least amount of information Alice and Bob need to exchange (i.e. reveal to each other) about their inputs to successfully perform T. This amount is expressed using mutual information (specifically, conditional mutual information (1.7)). We start by defining the *information cost* of a protocol π. Given a prior distribution μ on inputs (X, Y) the information cost

$$\mathsf{IC}(\pi, \mu) := I(Y; \Pi|X) + I(X; \Pi|Y), \tag{2.1}$$

where Π is the random variable representing a realization of the protocol's transcript, including the *public* randomness it used. In other words, (2.1) represents the sum of the amount of information Alice learns about Y by participating in the protocol and the amount of information Bob learns about X by participating. Note that the prior distribution μ may drastically affect $\mathsf{IC}(\pi, \mu)$. For example, if μ is a singleton distribution supported on one input (x_0, y_0), then $\mathsf{IC}(\pi, \mu) = 0$ for all π, since X and Y are already known to Bob and Alice respectively under the prior distribution μ. Definition (2.1), which will be justified shortly,

generalizes Shannon's entropy in the non-interactive regime. Indeed, in the transmission case, Bob has no input, thus $X \sim \mu$, $Y = \perp$, and Π allows Bob to reconstruct X, thus $\mathsf{IC}(\pi, \mu) = I(X; \Pi) = H(X) - H(X|\Pi) = H(X)$.

The *information complexity* of a task T can now be defined similarly to communication complexity in (1.15):

$$\mathsf{IC}(T, \mu) := \inf_{\pi \text{ successfully performs } T} \mathsf{IC}(\pi, \mu). \tag{2.2}$$

One notable distinction between (1.15) and (2.2) is that the latter takes an infimum instead of a minimum. This is because while the number of communication protocols of a given communication cost is finite, this is not true about information cost. One can have a sequence $\pi_1, \pi_2, \ldots$ of protocols of ever-increasing communication cost, but whose information complexity $\mathsf{IC}(\pi_n, \mu)$ converges to $\mathsf{IC}(T, \mu)$ in the limit. Moreover, as we will discuss later, this phenomenon is already observed in a task T as simple as computing the conjunction of two bits.

Our discussion of information complexity will be focused on the slightly simpler to reason about *distributional* setting, where inputs are distributed according to some prior μ. In (2.2), if T is the task of computing a function F with error ε w.r.t. μ, the distribution μ is used twice: first in the definition of "success", and then in measuring the amount of information learned. It turns out that it is possible to define worst-case information complexity [7] as the information complexity with respect to the worst-possible prior distribution in the spirit of the minimax relationship (1.16). In particular, the direct sum property of information complexity which we will discuss below holds for prior-free information complexity as well.

Information complexity as defined here has been extensively studied in a sequence of recent works [2, 6, 7, 12, 19, 28], and the study is still very much in progress. In particular, it is surprisingly simple to show that information complexity is additive for tasks over independent pairs of inputs. Let T_1 and T_2 be two tasks over pairs of inputs (X_1, Y_1), (X_2, Y_2), and let μ_1, μ_2 be distributions on pairs (X_1, Y_1) and (X_2, Y_2), respectively. Denote by $T_1 \otimes T_2$ to task composed of successfully performing both T_1 and T_2 on the respective inputs (X_1, Y_1) and (X_2, Y_2). Then information complexity is additive over these two tasks:

Theorem 2.1. $\mathsf{IC}(T_1 \otimes T_2, \mu_1 \times \mu_2) = \mathsf{IC}(T_1, \mu_1) + \mathsf{IC}(T_2, \mu_2)$.

Proof. (Sketch, a complete proof of a slightly more general statement can be found in [7]). The "easy" direction of this theorem is the '$\leq$' direction. Take two protocols π_1 and π_2 that perform T_1 and T_2 respectively, and consider the concatenation $\pi = (\pi_1, \pi_2)$ (which clearly performs $T_1 \otimes T_2$). Consider what Alice learns from an execution of π with prior $\mu_1 \times \mu_2$. A straightforward calculation using, for example, repeated application of the chain rule (1.8) yields

$$I(Y_1 Y_2; \Pi_1 \Pi_2 | X_1 X_2) = I(Y_1; \Pi_1 | X_1) + I(Y_2; \Pi_2 | X_2),$$

and similarly for what Bob learns. Therefore $\mathsf{IC}(\pi, \mu_1 \times \mu_2) = \mathsf{IC}(\pi_1, \mu_1) + \mathsf{IC}(\pi_2, \mu_2)$. By passing to the limit as $\mathsf{IC}(\pi_1, \mu_1) \to \mathsf{IC}(T_1, \mu_1)$ and $\mathsf{IC}(\pi_2, \mu_2) \to \mathsf{IC}(T_2, \mu_2)$ we obtain the '$\leq$' direction.

The '$\geq$' direction is more interesting, even if the proof is not much more complicated. In this direction we are given a protocol π for solving $T_1 \otimes T_2$ with information cost $I = \mathsf{IC}(\pi, \mu_1 \times \mu_2)$, and we need to construct out of it two protocols for T_1 and T_2 of information costs I_1 and I_2 that add up to $I_1 + I_2 \leq I$. We describe the protocol $\pi_1(X_1, Y_1)$ below:

- Bob samples a pair $(X_2, Y_2) \sim \mu_2$, and sends X_2 to Alice;
- Alice and Bob execute $\pi((X_1, X_2), (Y_1, Y_2))$, and output the portion relevant to T_1 in the performance of $T_1 \otimes T_2$.

It is not hard to see that the tuple (X_1, Y_1, X_2, Y_2) is distributed according to $\mu_1 \times \mu_2$, and hence by the assumption on π, π_1 successfully performs T_1. Note that there is a slight asymmetry in π_1: X_2 is known to both Alice and Bob while Y_2 is only known to Bob. For the purpose of correctness, the protocol would have worked the same if Bob also sent Y_2 to Alice, but it is not hard to give an example where the information cost of π_1 in that case is too high. The information cost of π is thus given by the sum of what Bob learns about X_1 from π_1 and what Alice learns about Y_1 (note that (X_2, Y_2) are not part of the input):

$$I_1 = I(X_1; \Pi | X_2 Y_1 Y_2) + I(Y_1; \Pi | X_1 X_2).$$

The protocol $\pi_2(X_2, Y_2)$ is defined similarly to π_1 in a skew symmetric way:

$\pi_{\mathbf{2}}(\mathbf{X_2}, \mathbf{Y_2})$:

- Alice samples a pair $(X_1, Y_1) \sim \mu_1$, and sends Y_1 to Bob;
- Alice and Bob execute $\pi((X_1, X_2), (Y_1, Y_2))$, and output the portion relevant to T_2 in the performance of $T_1 \otimes T_2$.

We get that π_2 again successfully performs T_2, and its information cost is:

$$I_2 = I(X_2; \Pi | Y_1 Y_2) + I(Y_2; \Pi | X_1 X_2 Y_1).$$

Putting I_1 and I_2 together we get:

$$\begin{aligned} I_1 + I_2 = I(X_1; \Pi | X_2 Y_1 Y_2) + I(Y_1; \Pi | X_1 X_2) + I(X_2; \Pi | Y_1 Y_2) + I(Y_2; \Pi | X_1 X_2 Y_1) = \\ I(X_2; \Pi | Y_1 Y_2) + I(X_1; \Pi | X_2 Y_1 Y_2) + I(Y_1; \Pi | X_1 X_2) + I(Y_2; \Pi | X_1 X_2 Y_1) = \\ I(X_1 X_2; \Pi | Y_1 Y_2) + I(Y_1 Y_2; \Pi | X_1 X_2) = I. \end{aligned}$$

Once again, passing to the limit, gives us the '$\geq$' direction, and completes the proof. □

If we denote an n-time repetition of a task T by $T^{\otimes n}$, then repeatedly applying Theorem 2.1 yields

$$\mathsf{IC}(T^{\otimes n}, \mu^n) = n \cdot \mathsf{IC}(T, \mu). \tag{2.3}$$

Thus information complexity is additive and has the *direct sum property*: the cost of n copies of T scales as n times the cost of one copy. This fact can be viewed as an extension of the property $H(X^n) = n \cdot H(X)$ to interactive problems, but what does it teach us about communication complexity?

Direct sum and interactive compression. Let us return to the communication complexity setting, fixing T to be the task of computing a function $F(X, Y)$ with some error at most $\varepsilon > 0$ over a distribution μ (the case $\varepsilon = 0$ seems to be different from $\varepsilon > 0$). We will denote by F_ε^n the task of computing n copies of F on independent inputs distributed according to μ^n, with error at most ε *on each copy* (note that computing F correctly with error at most ε on all copies simultaneously is a harder task). The *direct sum* question for communication complexity asks whether

$$D_{\mu^n}(F_\varepsilon^n) = \Omega(n \cdot D_\mu(F_\varepsilon))? \tag{2.4}$$

While this question remains open, information complexity sheds light on this question by linking it to problems in interactive coding theory. As discussed below, information complexity appears to be the best tool for either proving or disproving (2.4), as well as for establishing the "right" direct sum theorem in case (2.4) is false. It is an easy observation that the information cost of a protocol π is always bounded by its length $|\pi|$, and therefore information complexity is always bounded by communication complexity. Therefore, by (2.3),

$$\frac{1}{n} \cdot D_{\mu^n}(F_\varepsilon^n) \geq \frac{1}{n} \cdot \mathsf{IC}(F_\varepsilon^n, \mu^n) = \mathsf{IC}(F_\varepsilon, \mu). \tag{2.5}$$

It turns out that the converse is also true in the limit, as $n \to \infty$ [6]:

$$\lim_{n\to\infty} \frac{1}{n} \cdot D_{\mu^n}(F_\varepsilon^n) = \mathsf{IC}(F_\varepsilon, \mu). \tag{2.6}$$

Equation (2.6) can be viewed as the interactive version of the Source Coding Theorem (1.2). In particular, it gives an operational characterization of information complexity exclusively in terms of communication complexity.

A promising attack route (that works to-date followed) on the direct sum question for communication complexity is to try and prove a relationship of the type $\mathsf{IC}(F_\varepsilon, \mu) \gtrsim D_\mu(F_\varepsilon)$ (as discussed above, the converse is trivially true). Indeed, if we could prove that $\mathsf{IC}(F_\varepsilon, \mu) = \Omega(D_\mu(F_\varepsilon))$, by (2.5) it would imply that $\frac{1}{n} \cdot D_{\mu^n}(F_\varepsilon^n) = \Omega(D_\mu(F_\varepsilon))$ and prove (2.4).

One equivalent way to interpret the attempts to prove $\mathsf{IC}(F_\varepsilon, \mu) \gtrsim D_\mu(F_\varepsilon)$ is in terms of a search for an interactive analogue of Huffman coding (1.11) (where it does hold that $H(X) > C(x) - 1$). (One way) Huffman coding shows how to encode a low-entropy "uninformative" signal into a short one. Its interactive version seeks to simulate a low information cost "uninformative" protocol π with a low communication protocol π'.

Until very recently, we did not know whether such a general compression scheme exists. Just this year, the first example of a relation whose information and communication complexities are exponentially separated was given in a striking work by Ganor, Kol, and Raz [19]. This result, in particular, shows a protocol π for a sampling problem that has information cost I, but which cannot be simulated by a protocol π' with *communication* cost $< 2^{\Omega(I)}$.

Note that (2.5), which follows from Theorem 2.1, can be further sharpened as follows. If there is a protocol π_n for solving F_ε^n — n copies of F — with communication cost C_n, then there is a protocol π_1 for solving a single copy of F_ε whose communication cost is still at most $C := C_n$, and whose information cost is at most $I \leq C_n/n$. To prove a lower bound on C_n, we can assume that it is "too small", and then show how to convert π_1 into a protocol π' for F_ε that uses $< D_\mu(F_\varepsilon)$ communication. This brings us to the following general interactive coding/compression question:

Problem 2.2 (Interactive compression problem). *Given a protocol π whose communication cost is C and whose information cost is I, what is the smallest amount of communication needed to (approximately) simulate π?*

To prove the strongest possible direct sum theorem we need π' to be compressed all the way down to $O(I)$ bits of communication (the strongest possible interactive compression result), however, partial interactive compression results lead to weaker (but still non-trivial) direct sum theorems. At present, the two strongest compression results, which partially re-

solve Problem 2.2, compress π to $\tilde{O}(\sqrt{C \cdot I})$ communication[4] [2] and $2^{O(I)}$ communication [7], respectively. Note that these results are incomparable since $C > I$ can be much (e.g. double-exponentially) larger than I.

These result lead to direct sum theorems for randomized communication complexity. As the compression introduces an additional small amount of error, the first result implies for any constant $\rho > 0$:

$$D_{\mu^n}(F_\varepsilon^n) = \tilde{\Omega}(\sqrt{n} \cdot D_\mu(F_{\varepsilon+\rho})), \tag{2.7}$$

and the second one implies

$$D_{\mu^n}(F_\varepsilon^n) = \Omega(n \cdot \log(D_\mu(F_{\varepsilon+\rho}))). \tag{2.8}$$

The recent result of Ganor et al. [19] rules out the strongest possible direct sum theorem for relations. Since the hard-to-compress protocol in their example has a very high communication complexity (on the order of 2^{2^I}), it is still possible that any protocol can be compressed to $O(I \cdot \log^{O(1)}(C))$ communication, leading to a direct sum theorem with $\frac{n}{\log^{O(1)} n}$ instead of just n. We should also note that the direct sum situation with functions (as opposed to relations) remains open.

Why is interactive compression so much harder than non-interactive? The main difference between the interactive and non-interactive compression settings is that in the interactive setting each message of the protocol conveys an average of $I/C \ll 1$ bits of information. There are many ways to compress communication in the relevant setting, but all of them incur an average loss of $\Omega(1)$ bits per round (Huffman coding being one example of this phenomenon). This is prohibitively expensive in the interactive case, if the number of rounds of interaction r is equal to C. Therefore, inevitably, to compress interactive communication one has to compress multiple rounds in one message. This problem disappears when $I \gg r$, and this is what makes the '$\leq$' direction of (2.6) go through when n is sufficiently large.

Direct product for communication complexity. Next, we turn our attention to the more difficult *direct product* problem for communication complexity. The direct sum question talks about the amount of resources needed to achieve a certain probability of success on n copies of F. What if that amount of resources is not provided? For example, (2.5) implies that unless $n \cdot \mathsf{IC}(F_\varepsilon, \mu)$ bits of communication is allowed in the computation of F_ε^n, the computation of *some* copy of F will have $< 1 - \varepsilon$ success probability. What does it tell us about the success probability of *all* copies simultaneously? It only tells us that the probability of the protocol succeeding on all copies simultaneously is bounded by $1 - \varepsilon$. This is a very weak bound, since solving the n copies independently leads to a success probability of $(1 - \varepsilon)^n$, which is exponentially small for a constant ε. How can this gap be reconciled? In particular, can one show that Alice and Bob cannot "pool" the errors from all n copies on the same instances, thus keeping the success probability for each coordinate, as well as the global success probability, close to $1 - \varepsilon$? The direct product problem precisely addresses this question. Let us denote by $\mathsf{suc}(F, \mu, C)$ as the highest success probability (w.r.t. μ) in computing F that can be attained using communication $\leq C$. Thus $\mathsf{suc}(F, \mu, C) \geq 1 - \varepsilon$ is equivalent to $D_\mu(F_\varepsilon) \leq C$. Somewhat informally phrased, the direct product question asks whether

$$\mathsf{suc}(F^n, \mu^n, o(n \cdot C)) < \mathsf{suc}(F, \mu, C)^{\Omega(n)}? \tag{2.9}$$

[4]Here, the $\tilde{O}(\cdot)$ notation hides poly-logarithmic factors.

As with the direct sum question, the direct product question appears "obvious": one would expect that the best we can do is just execute the best protocol for one copy of F n times independently. This will lead to a success probability of $\leq \mathsf{suc}(F,\mu,o(C))^n$.

A prominent setting within complexity theory where a question similar to the direct product question arose is that of *parallel repetition* for two-prover games [39]. Parallel repetition is used in the context of probabilistically checkable proofs (PCP) and hardness amplification. Hardness amplification is accomplished here by taking a hard task T (e.g. a verification procedure where the success probability of an unauthorized provers is $1-\varepsilon$), and creating a task T^n by taking n independent instances of T. It has been shown [39] that as n grows, the success probability goes to 0. Unfortunately, it does not go to 0 as $(1-\varepsilon)^n$. Indeed, as shown by a counterexample constructed by Raz [40], the best rate one can hope for is $(1-\varepsilon^2)^n$. The reason for this, pointed out by an earlier example by Feige and Verbitsky [18], is that the answers can be arranged to align errors together, so that when the provers fail, they fail on a lot more than εn coordinates at the same time. This is possible when answers are allowed to be correlated.

The direct product question (2.9) for communication complexity combines features from the direct sum question (thus hinting that information complexity is to play a role here as well), and from the parallel repetition setup (since we want a success probability dropping exponentially in n). The direct sum discussion already suggests that for $\mathsf{suc}(F,\mu,C) = 1-\varepsilon$, the best scaling of the amount of communication one can hope for is as $n\cdot I$, where $I = \mathsf{IC}(F_\varepsilon,\mu)$. This is because, as $n\to\infty$, the per-copy communication cost of computing F with error ε scales as $n\cdot I$. Thus, if we denote by $\mathsf{suc}^{\mathsf{i}}(F,\mu,I) \geq \mathsf{suc}(F,\mu,I)$ the best success probability one can attain at solving F while incurring an *information cost* of at most I, the direct product question for information asks whether

$$\mathsf{suc}(F^n,\mu^n,o(n\cdot I)) < \mathsf{suc}^{\mathsf{i}}(F,\mu,I)^{\Omega(n)}? \tag{2.10}$$

Note that the success probability on the left-hand-side is still with respect to communication. A statement such as this with respect to information cost is bound to be false: Information cost being an average-case quantity, one can attain an information-cost I_n protocol by doing nothing with probability $1-\delta$, and incurring an information cost of $I_n/\delta \gg n\cdot I$ with probability δ that can be taken only *polynomially* (and not exponentially) small.

In a sequence of two papers, the second being very recent [11, 12], (2.10) was shown to be true up to polylogarithmic factors for boolean functions. To simplify parameters, suppose $\mathsf{suc}^{\mathsf{i}}(F,\mu,I) < 2/3$. Then there are constants c_1, c_2 such that

$$\text{if } T\log T < c_1 n\cdot I, \text{ then } \mathsf{suc}(F^n,\mu^n,T) < 2^{-c_2 n}. \tag{2.11}$$

The proof of (2.11) is quite involved and combines ideas from the proof of direct sum theorems and of parallel repetition theorems.

Exact communication complexity bounds. One of the great successes of information theory as it applies to (classical, one-way) communication problems is in its ability to give precise answers to fairly complicated asymptotic communication problems, for example ones involving complicated dependencies between terminals or complicated channels. For example, the capacity of the binary symmetric channel $BSC_{0.2}$ is precisely $1-H(0.2)\approx 0.278$, which means that to transmit n bits over such a channel, we will need $\approx 3.596n$ utilizations of the channel (i.e. will need to send $\approx 3.596n$ bits down the channel). Using combinatorial

techniques, in most cases, such precision is inaccessible in the two-party setting, since the techniques often lose constant factors by design. In contrast, information complexity extends the precision benefits of one-way information theory to the interactive setting.

We give one specific example of an exact communication complexity bound. Recall that the disjointness problem $Disj_n(X,Y)$ takes two n-bit vectors X,Y and checks whether there is a location with $X_i = Y_i = 1$. Thus $Disj_n$ is just a disjunction of n independent copies of the two bit $AND(X_i,Y_i)$ function. Using techniques similar to the proof of Theorem 2.1, one can show that the communication complexity of disjointness is tightly linked with the information complexity of AND. Note that disjointness becomes trivial if many coordinates (X_j,Y_j) of the input are $(1,1)$. However, any distribution of inputs where $\mu((X_j,Y_j)=(1,1)) \sim 1/n \to 0$ will not be trivial. More formally, denote by 0^+ a function $f(n)$ of n such that $f(n) = o(1)$ and $f(n) \gg 2^{-O(n)}$. For example, one can take $f(n) = 1/n$. Then with some work one shows [9] that

$$R_{0^+}(Disj_n) = \left(\inf_{\mu:\mu(1,1)=0} \mathsf{IC}(AND_0,\mu)\right)\cdot n \pm o(n). \tag{2.12}$$

Thus, understanding the precise asymptotics of the communication complexity of $Disj_n$ boils down to understanding the (0-error) information complexity of the two-bit AND function[5]. It turns out that one can give an explicit information-theoretically optimal family of protocols for AND, and calculate the quantity in (2.12) explicitly, obtaining $R_{0^+}(Disj_n) = C_{DISJ}\cdot n \pm o(n)$ where $C_{DISJ} \approx 0.4827$.

Interestingly, even in the case of such a simple function as two-bit AND, the information complexity is not attained by any particular protocol, but rather by an infinite family of communication protocols! Moreover, if we denote by $IC_r(AND_0)$ the information cost of AND where the infimum in (2.2) is only taken over protocols of length r, then it turns out that $IC_r(AND_0) = IC(AND_0) + \Theta(1/r^2)$, implying that an asymptotically optimal protocol is only achieved with a super-constant number of rounds [9]. We do not yet know how general this $1/r^2$ gap phenomenon is, and which communication tasks admit a minimum in (2.2).

3. Interactive error-correcting codes

Adversarial error-correction. The discussion so far focused on coding for interactive computing over a noiseless binary channel. In this section we will focus on error-correction problems when the channel contains random or adversarial noise. The first regime we would like to consider is that of adversarial noise. In this regime Alice and Bob are trying to perform a task T over a channel in which an adversary is allowed to corrupt a constant fraction of the messages. Both the regime of a binary channel and that of a channel with constant-size alphabet Σ (i.e. where symbols $\sigma \in \Sigma$ are being transmitted over the channel) are interesting.

The one-way case has been extensively studied for several decades, as discussed in the introduction. If the task T is just a simple transmission task, then the theory of (worst-case) error-correcting codes [34, 44] applies. While there are many open problems in coding theory, the overall picture is fairly well understood. In particular, constructions of "good"

[5]Note that even when $\mu(1,1) = 0$ and thus $AND(X,Y) = 0$ on supp(μ), the task AND_0 requires the protocol to *always* be correct – even on the $(1,1)$ input. Otherwise, $\mathsf{IC}(AND_0,\mu)$ would trivially be 0.

positive-rate, constant-distance codes exist (i.e. codes that increase communication by a constant factor only, and can tolerate a constant fraction of errors), and there are efficient encoding and decoding constructions.

In the interactive case, the task may include many back-and-forth messages. As a generic task, it is convenient to think about alternating binary pointer jumping (BPJ_n). In this problem the parties are working with a depth-n binary tree. Alice is given a subset T_A of edges on the odd layers of the tree, with exactly one edge coming out of each vertex on odd layers. Similarly, Bob is given a subset T_B of edges on the even layers of the tree. Their goal is to find the unique leaf that is connected to the root by edges from $T_A \cup T_B$. There is an obvious n-bit protocol for finding the leaf, where Alice and Bob alternate. The definition of BJP_n is parallel to the definition of a n-round protocol π as given by Definition 1.1. In this sense, BPJ_n is the generic interactive task, as any interactive protocol can be recast as an instance of BPJ_n.

To continue the comparison with the non-interactive setting, suppose an adversary is allowed to corrupt a δ-fraction of the symbols exchanged by Alice and Bob, for some $\delta > 0$. Can they still compute BPJ_n? Solving BPJ_n efficiently requires a lot of back-and-forth interaction. A naïve approach would be to apply (standard) error-correction to the interactive protocol on a round-by-round basis. This does not work, because the adversary can concentrate all of her errors, for example, on the first round, causing all subsequent communications to be wrong and derailing the protocol's execution. Another obvious solution that dose work is to have Alice send her input T_A to Bob using a standard error-correcting code. Bob then can compute the leaf. This solution works, but causes an exponential blow-up in communication, since T_A takes $\sim 2^n$ bits to describe, while the efficient solution for BPJ_n requires only $O(n)$ communication.

It is not at all clear that a constant-rate error correcting code is possible. Surprisingly, constant-rate error-correcting codes for interactive computing do exist. The first such code was demonstrated in a breakthrough work by Schulman in the 1990s [42], who showed a constant-rate code against an adversary who is allowed to corrupt a constant δ-fraction of the symbols on the channel for $\delta < 1/240$. Schulman introduced a concept of a *tree code*. Variants and extensions of tree codes have been used in all constructions since. The construction opened up opportunities for interactive error-correction, but also left room for improvement, as the error-parameter $\delta < 1/240$ is far from optimal and the error-correction is not efficient in that it requires time exponential in n to compute the encoding/decoding (even though the communication itself is $O(n)$ symbols).

After a gap in progress on interactive error-correction, a substantial amount of progress has been made in the last 5 years [5, 8, 10, 20–22]. Progress so far has focused on (1) making the tolerable error rate δ as high as possible; (2) making the construction explicit and computationally efficient. This while keeping the rate (i.e. the ratio between the encoding length and the length of the noise-free execution) of the code constant. What remains completely open is the exact coding rate for interactive coding, given a specific value of δ. All we know are characterizations of δ for which various specific types of good codes exist.

Next, let us discuss the error-rate region for which (two-party) interactive error-correction is possible. Suppose Alice and Bob communicate over a channel which uses an alphabet Σ_2 with $|\Sigma_2| = O(1)$ a large constant that is allowed to depend on δ (the case of a binary noisy channel, $|\Sigma_2| = 2$, is also interesting, with many of the problems still open there). An interactive error-correction scheme π is a protocol of a fixed length $n' = O(n)$ over Σ_2 that solves BPJ_n, even when the channel is affected by a noise of rate δ. In other words, for

any inputs (T_A, T_B), any execution transcript Π of π in which a total of at most $\delta \cdot n'$ of the symbols were corrupted results with Alice and Bob outputting the correct leaf

$$BPJ_n(T_A, T_B) = D_A(T_A, \Pi) = D_B(T_B, \Pi), \tag{3.1}$$

D_A and D_B being the decoding functions for Alice and Bob, respectively. Here D_A and D_B are only allowed to depend on the portions of the transcript Π accessible to Alice and Bob, respectively.

First assume that in π, the player speaking in each round is pre-determined (a single symbol is sent in each round). Such protocols are caller *robust*. Note that without this assumption, it is possible to have a round in which both Alice and Bob (or neither Alice nor Bob) speak, since error may confuse the players as to whose turn it is to speak. In this case further modeling assumptions are needed to specify what happens during these rounds.

In the robust case, note that the adversary knows ahead of time n_A and n_B — the number of rounds Alice and Bob speak, respectively, in π. Here $n_A + n_B = n'$. Assume without loss of generality that $n_A \leq n'/2$. Then, as with the proof that one way error-correcting codes cannot recover from an error rate exceeding $1/2$, by extrapolating between $\pi(T_{A_1}, T_B)$ and $\pi(T_{A_2}, T_B)$, an adversary can corrupt $n_A/2$ rounds of π, and prevent Bob from distinguishing two potential inputs T_{A_1} and T_{A_2} of Alice. If the resulting transcript is Π, as long as $BPJ_n(T_{A_1}, T_B) \neq BPJ_n(T_{A_2}, T_B)$, either $D_B(T_B, \Pi) \neq BPJ_n(T_{A_1}, T_B)$ or $D_B(T_B, \Pi) \neq BPJ_n(T_{A_2}, T_B)$, meaning that π sometimes fails. Thus the adversary can foil the protocol using $n_A/2 \leq n'/4$ errors, so we cannot hope to overcome an error rate of $\delta \geq 1/4$.

It turns out [10] that it is possible to deal with error rates of $\delta = 1/4 - \varepsilon$ using constant-rate codes. As in Schulman's construction, the key technical ingredient of this result is that of a *tree code*. A tree code is a prefix code $C : \{0,1\}^m \to \Sigma_2^m$; in a prefix code the i-th symbol of the codeword $C(S)_i = C_i(S_{[1..i]})$ only depends on the first i symbols of the word being encoded. It is clear that a prefix code cannot have the constant-distance property since, for example $C(0^m)$ and $C(0^{m-1}1)$ cannot differ in more than one symbol. The best property we can hope for is that codewords of length k that deviate after the i-th symbol will differ by close to $(k - i)$ symbols. This is indeed the definition of a tree code: a tree code $C : \{0,1\}^m \to \Sigma_2^m$ is said to have *distance* α if for all i,k, and $w \in \{0,1\}^i$, $w_0, w_1 \in \{0,1\}^{k-i-1}$,

$$d_H(C(w0w_0), C(w1w_1)) \geq \alpha \cdot (k - i). \tag{3.2}$$

It can be shown [42] that tree codes exist for any constant $\alpha < 1$ (the alphabet Σ_2 may need to be made sufficiently large, with its size increasing as α approaches 1). Note that it is easy to see that a random code will *not* be a tree code with a very high probability. Therefore, even constructing a non-explicit tree codes is not a trivial task. To decode a tree code, the receiver just finds the codeword that is closest to the received word in Hamming distance.

Informally, each symbol sent by the tree code not only encodes the current symbol being sent, but also hashes the entire history of the transmission, ensuring that a mistake introduced by an adversary will be corrected as following rounds arrive. The key useful property of tree codes for the purposes of interactive error-correction codes is the following: Suppose that t rounds ago Alice sent a message z encoded using the tree code, and the adversary managed to keep Bob from receiving it, and instead Bob thinks that $\bar{z}$ was sent t rounds ago. This means that the amount of errors between now and some point $t' \geq t$ rounds ago must be at

least $\alpha \cdot t'/2$. In other words, to keep Bob from learning z, the adversary has to introduce many errors in a large stretch that in particular is at least t symbols long.

Next, let us give the intuition for how tree codes can be useful in interactive error correction, following the construction in [10]. Unfortunately, due to space constraints, we will not be able to give a full sketch here. The protocol π will proceed by having Alice and Bob send edges of T_A and T_B, respectively, using a tree code to encode a stream of edges being sent. The parties are trying to build the unique path from the root in $T_A \cup T_B$. At each point in time, one of the parties (say Alice) can extend the path, assuming she correctly decoded the previous edges. By the discussion above about the main property of tree codes, to keep Alice from correctly decoding the previous edges, the adversary will have to use an error rate of at least $\alpha/2$ in Bob's transmissions between the time the previous edge had been sent by Bob, and when it is decoded by Alice. This amounts to an error rate of $\alpha/4$. By choosing $\alpha/4 > \delta$ (which is possible since $\delta < 1/4$) we can guarantee enough rounds in which Alice and Bob will make progress. This outline glosses over how edges are represented, and indeed representing edges so that each only takes $O(1)$ bits which can be encoded using the tree code is the main technical challenge overcome by [10].

As noted earlier, relaxing the robustness assumption requires further modeling assumptions on what happens in rounds where either both Alice and Bob or neither speak. One would expect that by having the party that is being targeted by the adversary speak more, one can improve the error tolerance of the protocol. Indeed, the example showing the $1/4$ limit above could be remedied if the party being targeted by the adversary spoke more than $n'/2$ of the rounds (thus forcing the adversary to expend more of her budget). Under a reasonable model, a recent work [22] shows that the error-tolerance of non-robust protocols can be made $2/7 - \varepsilon > 1/4$, and that this bound is tight.

In the one-way error-correcting coding theory, an important way of going beyond error-rate $1/2$ is using the concept of *list decoding*. A list-decodable code is one where for a corrupt encoded words, there is a (constant-size) list of possible decodings. Over large, constant-size alphabets, list-decodable codes exist for any error rate of $1-\varepsilon$, where the output list size is $O_\varepsilon(1)$. In the interactive setting, somewhat surprisingly, one can also construct list-decodable error-correcting schemes. In the robust setting, the best error rate attainable by a constant-rate code is $1/2 - \varepsilon$ [8]. This construction uses a generalization of tree codes called list-tree-codes. This generalization has an average-case rather than worst-case coding property, and is instantiated by a random prefix code with a sufficiently large constant $|\Sigma_2|$ with a very high probability. Interestingly, it appears that one needs interactive list-decoding even just to attain optimal error-resistance for unique decoding in some regimes.

One limitation of the constructions above is that they are not explicit. In other words, while we know that they can all be instantiated, often with a random prefix code, no provable explicit constructions of tree codes and list-tree-codes are known. Worse yet, even if one could somehow derandomize these constructions, the brute-force decoding procedures require exponential time. Several recent works developed efficient interactive error-correcting coding schemes. In particular, the very recent work by Ghaffari and Haeupler [21] gives an efficient scheme that achieves the same error-correction guarantees as the best-known non-efficient scheme (see [21] for additional recent history and references). Its only limitation is that it uses randomness for initialization, but it allows this randomness to be accessible by the adversary, so it is not a major limitation since no shared secret between Alice and Bob is needed. Most excitingly, while the scheme has a slightly sub-constant rate, by combining it with the construction of [8] it appears that it can be made constant-rate, thus concluding the

quest for efficient interactive error-correcting schemes with optimal error dependence.

All efficient schemes to-date follow a similar paradigm: start with a non-efficient scheme on a very small scale (say, $\log\log n$ rounds). On such a small scale one can just brute-force the search for tree codes, and for efficient encoding-decoding schemes. Next, show how to go from an interactive constant-rate error-correcting scheme of depth k to one of depth, say, 2^k. Note that one will only need to apply such a transition twice to go from depth $\log\log n$ do depth n.

A major gap in our understanding of interactive error-correction is in the *rate* of optimal codes. In other words, for a given error rate $\delta = 1/4 - \varepsilon$, what is the best rate

$$\rho_\delta = \frac{n}{n' \log |\Sigma_2|}$$

one can hope to attain in solving BPJ_n? We do not even know the asymptotics of ρ_δ as δ approaches the boundary points of 0 and $1/4$. Perhaps this should not be too disappointing, since parallel questions are open for one-way communication. However, one could hope to resolve these problems in the *random error* model, since there Shannon's classical work does give us precise channel capacity answers. We turn our attention to that regime next.

Random errors and channel capacity. In the random error model, Alice and Bob communicate over a noisy channel $\mathcal{C}$, where the noise is generated randomly. For concreteness, we will focus here on the binary symmetric channel with error ε, BSC_ε, where bits are being transmitted and each bit sent over the channel is independently flipped with probability ε.

As discussed earlier, the channel capacity of BSC_ε is given by (1.10) and is equal to $1 - H(\varepsilon)$. Informally, this means that the utility of BSC_ε in conducting communication is $1 - H(\varepsilon)$, and that for a growing n, transmitting n random bits over BSC_ε will require

$$n/\mathsf{cap}(BSC_\varepsilon) \pm o(n) = n/(1 - H(\varepsilon)) \pm o(n) \tag{3.3}$$

utilizations of the channel. How this logic should extend to the interactive case is still up to debate. One natural extension is to consider the pointer jumping problem BPJ_n from before as the standard interactive problem, and to define interactive channel capacity in terms of the number of channel utilizations needed to execute BPJ_n, similarly to

$$\mathsf{icap}_1(\mathcal{C}) := \lim_{n\to\infty} \frac{n}{\#\text{ of utilizations of } \mathcal{C} \text{ needed to perform } BPJ_n \text{ w.h.p.}}. \tag{3.4}$$

No explicit formulas (or ways of obtaining explicit values) of $\mathsf{icap}_1(BSC_\varepsilon)$ are known. Even establishing directly that $\mathsf{icap}_1(BSC_\varepsilon) > 0$ does not seem completely straightforward, although this fact is a direct consequence of the more general adversarial setting from the previous section. One important recent result by Kol and Raz [29] establishes a gap between icap_1 and Shannon's channel capacity for BSC_ε showing that

$$\mathsf{icap}_1(BSC_\varepsilon) = 1 - \Theta(\sqrt{H(\varepsilon)}) = 1 - \Theta(\sqrt{\varepsilon \log 1/\varepsilon}) < \mathsf{cap}(BSC_\varepsilon) \tag{3.5}$$

as $\varepsilon \to 0$. This result is quite technical, and underscores the difficulty of the interactive channel capacity question.

One of the nice properties of Shannon's one-way information theory is that the notions of entropy and of channel capacity commute. That is, if we want to transmit a random variable

X whose entropy is $H(X) \gg 1$ over a channel $\mathcal{C}$, then the number of channel utilizations needed to transmit X is on average

$$(H(X)/\mathsf{cap}(\mathcal{C}))(1 + o(1)). \tag{3.6}$$

In the interactive setting, we have established information complexity as the interactive analogue of channel capacity. It is unclear whether there is a way to define interactive channel capacity that makes the interactive analogue of (3.6) hold. Such an analogue may also help shed light onto the basic structure of interactive communication. The result of [29] implies that such a characterization cannot simultaneously capture interactive and non-interactive tasks, and thus it is bound to be quite complex.

4. Conclusion and discussion

We conclude with some specific open problems and a general discussion. In addition to some of the open questions outlined above, particularly surrounding compressibility of interactive computation, several other questions, that are easy in the non-interactive setting arise when interaction is added to the mix.

Computability of information complexity. The first problem that is (somewhat embarrassingly) open, is computing the information complexity from the truth table of F:

Problem 4.1. *Given the truth table of a function $F : (X, Y) \mapsto \{0, 1\}$, and error parameter $\varepsilon \geq 0$, and a distribution μ of (X, Y), can one give a general procedure for computing the information complexity $\mathsf{IC}(F_\varepsilon, \mu)$?*

We believe the answer to Problem 4.1 to be affirmative. As noted above, the problem is that there might be a sequence of protocols whose information cost decreases as protocol size increases. The $\leq$ direction of (2.6) gives one way to obtain a decreasing sequence that converges to $\mathsf{IC}(F_\varepsilon, \mu)$ by considering the amortized cost of n copies of F as $n \to \infty$. Unfortunately, for this procedure to compute $\mathsf{IC}(F_\varepsilon, \mu)$, we need to have an effective bound on the sequence's rate of convergence down to $\mathsf{IC}(F_\varepsilon, \mu)$. The work [32] gives a computable characterization of $\mathsf{IC}(F_\varepsilon, \mu)$, but only when one fixes the number of rounds of interaction (back-and-forth messages) in advance. Once again, we do not know an effective rate of convergence of the round-restricted information complexity to the unrestricted value.

One can also formulate Problem 4.1 as a continuous dynamic programming problem in the spirit of the Hamilton-Jacobi-Bellman equation [9], but it is not clear how to solve the resulting equation, although it might be doable by better understanding the properties of the function $\mathsf{IC}(F_\varepsilon, \mu)$ when considered as a function on the space of distributions μ.

Multi-party communication. It is a natural and very interesting goal to generalize the discussion above to more than two terminals. There are various models for multi-terminal interactive computation. The main complication stems from the fact that the prior distributions, and the way the inputs to different players are correlated, may be rather sophisticated. One popular model of multi-party computation is that of number-on-forehead (NOF). In the NOF model each party gets to see all inputs but its own and the goal is to compute a function

$F(X_1, \ldots, X_k)$ of the inputs [14, 30]. Lower bounds in this model would have profound implications in complexity theory [3]. Multiparty NOF lower bounds are considerably harder than two-party bounds. For example, it is still unknown whether the communication complexity of the 3-party analogue of $Disj_n$ has communication complexity $\Theta(\sqrt{n})$ or $\Theta(n)$ (or something in between) [43].

There are numerous complications in extending notions of information complexity to multi-terminal settings. Apart from sheer technical difficulties, a major obstacle is finding the "right" analogue of public and private randomness. Note that even with three parties we have seven different types of randomness (one "private" for each party, one "public", and three shared between two of the three parties but not the third). Allowing all the different types of randomness leads to another impasse, as in this regime there are information-theoretically secure protocols for multi-party computation [4] which would bring the information complexity of all problems close to 0.

Beyond communication: continuous relaxations for other models of computation. From the viewpoint of theoretical computer science, information complexity can be viewed as the continuous relaxation of communication complexity. Avoiding the "discreteness" of bits and switching to information instead simplified not only the proofs, but the results themselves. For example, the direct sum theorem (Theorem 2.1) is true for information complexity but is not true, at least in full generality, for communication complexity. Thus, this is one more example in the context of complexity theory where a continuous relaxation is easier to deal with. There are many more such examples in the context of algorithms. For example, one of the leading paradigms in approximation algorithms involves relaxing discrete problems into continuous convex optimization programs (for example, linear or semi-definite), and then rounding the resulting fractional solution to obtain an integral one. This allows one to connect the problem of algorithm development with a rich (and deep) theory of continuous analysis and geometry.

In the context of computational complexity, there is still much to be desired in terms of our ability to "de-discretize" computation. The difficulty of dealing with a discrete computation theory has been foreseen by von Neumann as early as 1948 [46] in his Hixon Symposium talk:

> "There exists today a very elaborate system of formal logic, and, specifically, of logic as applied to mathematics. This is a discipline with many good sides, but also with certain serious weaknesses. This is not the occasion to enlarge upon the good sides, which I have certainly no intention to belittle. About the inadequacies, however, this may be said: Everybody who has worked in formal logic will confirm that it is one of the technically most refractory parts of mathematics. The reason for this is that it deals with rigid, all-or-none concepts, and has very little contact with the continuous concept of the real or of the complex number, that is, with mathematical analysis. Yet analysis is the technically most successful and best-elaborated part of mathematics. Thus formal logic is, by the nature of its approach, cut off from the best cultivated portions of mathematics, and forced onto the most difficult part of the mathematical terrain, into combinatorics.
>
> The theory of automata, of the digital, all-or-none type, as discussed up to now, is certainly a chapter in formal logic. It would, therefore, seem that it will have

to share this unattractive property of formal logic. It will have to be, from the mathematical point of view, combinatorial rather than analytical."

Over 65 years later, most fundamental problems in the theory of computation such as the $\mathbf{P}$ vs. $\mathbf{NP}$ problem are wide open, and most unconditional lower bounds are based on diagonalization ideas of Cantor, Gödel and Turing. Von Neuman's prognostication appears to have withstood the test of time.

Is there a natural continuous relaxation of computational complexity specific enough to deal with its major open problems? And are our mathematical tools mature enough to pursue one if it exists? It is hard to know, but information theory is a great example of a continuous theory that organizes (and greatly simplifies) discrete communication. Communication complexity started out as a discrete theory, but appears to be amenable to continuous treatment, with information complexity being its natural continuous relaxation. It will be very interesting to see whether this push can be extended further into computational complexity.

Acknowledgments. The author's work is supported in part by an NSF CAREER award (CCF-1149888), NSF CCF-0832797, NSF CCF-1215990, a Turing Centenary Fellowship, and a Packard Fellowship in Science and Engineering. I would like to thank Ankit Garg, Rotem Oshman, Denis Pankratov, and Omri Weinstein for their numerous comments on earlier drafts of this paper.

References

[1] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar, *An information statistics approach to data stream and communication complexity*, Journal of Computer and System Sciences **68** (2004), no. 4, 702–732.

[2] B. Barak, M. Braverman, X. Chen, and A. Rao, *How to compress interactive communication*, SIAM Journal on Computing **42** (2013), no. 3, 1327–1363.

[3] Richard Beigel and Jun Tarui, *On acc [circuit complexity]*, Foundations of Computer Science, 1991. Proceedings., 32nd Annual Symposium on, IEEE, 1991, pp. 783–792.

[4] Michael Ben-Or, Shafi Goldwasser, Joe Kilian, and Avi Wigderson, *Multi-prover interactive proofs: How to remove intractability assumptions*, Proceedings of the 20th Annual ACM Symposium on Theory of Computing, 1988, pp. 113–131.

[5] Z. Brakerski and Y.T. Kalai, *Efficient interactive coding against adversarial noise*, Electronic Colloquium on Computational Complexity (ECCC), 2012.

[6] M. Braverman and A. Rao, *Information equals amortized communication*, 52nd Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2011, pp. 748–757.

[7] Mark Braverman, *Interactive information complexity*, Proceedings of the 44th symposium on Theory of Computing, ACM, 2012, pp. 505–524.

[8] Mark Braverman and Klim Efremenko, *List and unique coding for interactive communication in the presence of adversarial noise*, Electronic Colloquium on Computational Complexity (ECCC) (2014).

[9] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein, *From information to exact communication*, Proceedings of the 45th annual ACM symposium on Symposium on theory of computing, ACM, 2013, pp. 151–160.

[10] Mark Braverman and Anup Rao, *Towards coding for maximum errors in interactive communication*, Proceedings of the 43rd annual ACM symposium on Theory of computing, ACM, 2011, pp. 159–166.

[11] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff, *Direct products in communication complexity*, Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on, IEEE, 2013, pp. 746–755.

[12] Mark Braverman and Omri Weinstein, *An interactive information odometer with applications*, Electronic Colloquium on Computational Complexity (ECCC), 2014.

[13] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Yao, *Informational complexity and the direct sum problem for simultaneous message complexity*, Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (Los Alamitos, CA) (Bob Werner, ed.), IEEE Computer Society, October 14–17 2001, pp. 270–278.

[14] Arkadev Chattopadhyay and Toniann Pitassi, *The story of set disjointness*, ACM SIGACT News **41** (2010), no. 3, 59–85.

[15] Thomas M Cover and Joy A Thomas, *Elements of information theory, 2nd edition*, J. Wiley and Sons, New York, 2006.

[16] Peter Elias, *List decoding for noisy channels*, (1957).

[17] Tomas Feder, Eyal Kushilevitz, Moni Naor, and Noam Nisan, *Amortized communication complexity*, SIAM Journal on Computing **24** (1995), no. 4, 736–750.

[18] Uriel Feige and Oleg Verbitsky, *Error Reduction by Parallel Repetition–A Negative Result*, Combinatorica **22** (2002), no. 4, 461–478.

[19] Anat Ganor, Gillat Kol, and Ran Raz, *Exponential separation of information and communication*, Electronic Colloquium on Computational Complexity (ECCC), 2014.

[20] R. Gelles, A. Moitra, and A. Sahai, *Efficient and explicit coding for interactive communication*, Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on, IEEE, 2011, pp. 768–777.

[21] Mohsen Ghaffari and Bernhard Haeupler, *Optimal error rates for interactive coding ii: Efficiency and list decoding*, arXiv preprint arXiv:1312.1763 (2013).

[22] Mohsen Ghaffari, Bernhard Haeupler, and Madhu Sudan, *Optimal error rates for interactive coding i: Adaptivity and other settings*, arXiv preprint arXiv:1312.1764 (2013).

[23] Venkatesan Guruswami, *List decoding of error-correcting codes*, Springer, 2004.

[24] ———, *Bridging shannon and hamming: List error-correction with optimal rate*, Proceedings of ICM, 2010.

[25] David A Huffman et al., *A method for the construction of minimum redundancy codes*, Proceedings of the IRE **40** (1952), no. 9, 1098–1101.

[26] Bala Kalyanasundaram and Georg Schnitger, *The probabilistic communication complexity of set intersection*, SIAM Journal on Discrete Mathematics **5** (1992), no. 4, 545–557. MR 93j:68080

[27] Mauricio Karchmer, Ran Raz, and Avi Wigderson, *Super-logarithmic depth lower bounds via the direct sum in communication complexity*, Computational Complexity **5** (1995), no. 3/4, 191–204, Prelim version CCC 1991.

[28] Iordanis Kerenidis, Sophie Laplante, Virginie Lerays, Jérémie Roland, and David Xiao, *Lower bounds on information complexity via zero-communication protocols and applications*, Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on, IEEE, 2012, pp. 500–509.

[29] Gillat Kol and Ran Raz, *Interactive channel capacity*, Proceedings of the 45th annual ACM symposium on Symposium on theory of computing, ACM, 2013, pp. 715–724.

[30] Eyal Kushilevitz and Noam Nisan, *Communication complexity*, Cambridge University Press, Cambridge, 1997.

[31] Troy Lee and Adi Shraibman, *Lower bounds in communication complexity*, Now Publishers Inc, 2009.

[32] N. Ma and P. Ishwar, *Some results on distributed source coding for interactive function computation*, Information Theory, IEEE Transactions on **57** (2011), no. 9, 6180–6195.

[33] _____, *The infinite-message limit of two-terminal interactive source coding*, Information Theory, IEEE Transactions on **59** (2013), no. 7, 4071–4094.

[34] F. J. MacWilliams and N. J. A. Sloane, *The theory of error correcting codes*, North–Holland, New York, 1977.

[35] Ilan Newman, *Private vs. common random bits in communication complexity*, Information Processing Letters **39** (1991), no. 2, 67–71.

[36] A Orlitsky and A El Gamal, *Communication complexity*, Complexity in information theory, Springer, 1988, pp. 16–61.

[37] Alon Orlitsky and James R Roche, *Coding for computing*, Information Theory, 1995. Proceedings., 1995 IEEE International Symposium on, IEEE, 1995, p. 451.

[38] Vera Pless, Richard A Brualdi, and William Cary Huffman, *Handbook of coding theory*, Elsevier Science Inc., 1998.

[39] R. Raz, *A parallel repetition theorem*, SIAM Journal on Computing **27** (1998), no. 3, 763–803.

[40] _____, *A counterexample to strong parallel repetition*, SIAM Journal on Computing **40** (2011), no. 3, 771–777.

[41] Alexander Razborov, *On the distributed complexity of disjointness*, TCS: Theoretical Computer Science **106** (1992).

[42] Leonard J. Schulman, *Coding for interactive communication*, IEEE Transactions on Information Theory **42** (1996), no. 6, 1745–1756.

[43] Alexander A Sherstov, *Communication lower bounds using directional derivatives*, Proceedings of the 45th annual ACM symposium on Symposium on theory of computing, ACM, 2013, pp. 921–930.

[44] M. Sudan, *Algorithmic introduction to coding theory – course notes*, 2001, http://people.csail.mit.edu/madhu/FT01/course.html.

[45] Jacobus Hendricus Van Lint, *Introduction to coding theory*, vol. 86, Springer, 1982.

[46] J. von Neumann, *The general and logical theory of automata*, John von Neumann, collected works, Pergamon Press, 1951, pp. 288–328.

[47] John M Wozencraft, *List decoding*, Quarterly Progress Report **48** (1958), 90–95.

[48] Andrew C. C. Yao, *Some complexity questions related to distributive computing (preliminary report)*, Proceedings of the eleventh annual ACM symposium on Theory of computing, ACM, 1979, pp. 209–213.

[49] _____, *Lower bounds by probabilistic arguments*, Foundations of Computer Science, 1983., 24th Annual Symposium on, IEEE, 1983, pp. 420–428.

Department of Computer Science, Princeton University, Princeton, NJ 08544, USA
E-mail: mbraverm@cs.princeton.edu

Counting constraint satisfaction problems

Andrei A. Bulatov

Abstract. Counting constraint satisfaction problems (CSPs) originate from two very different areas: statistical physics, where partition functions appearing in "spin-glass" models have been studied since the beginning of the last century, and counting combinatorial problems formally introduced by Valiant in the late 70s. In spite of such a long history, the systematic study of the general counting CSP started less than 15 years ago. In this short survey we review recent results on counting CSPs.

Mathematics Subject Classification (2010). Primary 68Q25; Secondary 68Q17.

Keywords. constraint satisfaction problem, counting, complexity, partition function, homomorphism.

1. Introduction

In a counting problem the aim is to find the number of certain arrangements in a given combinatorial structure. Counting problems and their generalizations frequently occur in a variety of areas ranging from combinatorics, to computer science, to statistical physics. In complexity theory a systematic study of counting problems was initiated by Valiant who in an attempt to understand the hardness of computing the Permanent introduced a formal framework for such problems [61, 62]. In particular, Valiant introduced the complexity class #P, as the class of non-negative integer functions $f(I)$, for which there is a nondeterministic Turing machine having $f(I)$ accepting paths when run on instance I; he also introduced reduction between counting problems, and showed the #P-hardness of several problems including the Permanent. Further research on counting problems, leaving aside determining the complexity of particular problems, included clarifying relations of #P to other complexity classes [1, 3, 38, 41, 57], establishing a hierarchy of counting complexity classes inside #P [4, 59], and the descriptive complexity of counting problems [55].

In this short survey we consider a special albeit quite broad type of counting problems, the counting version of the Constraint Satisfaction Problem (CSP). The CSP provides a powerful framework to express many combinatorial problems in a uniform way. In an instance of the CSP we are given a set D, a set of variables V, and a conjunctive formula $R_1(\mathbf{x}_1) \wedge \ldots \wedge R_m(\mathbf{x}_m)$, where each R_i is a relation (a predicate) on D, and each $\mathbf{x}_i$ is a tuple of variables from V, whose length matches the arity of R_i. The goal is to decide whether there is an assignment $\sigma : V \to D$ that satisfies the formula. Feder and Vardi [34] observed that the CSP can also be reformulated as the homomorphism problem: Given relational structures, $\mathcal{G}$ and $\mathcal{H}$, of the same signature, decide whether or not there exists a homomorphism from $\mathcal{G}$ to $\mathcal{H}$. In the counting version of the CSP, #CSP, the goal is to find the number of

Proceedings of the International Congress of Mathematicians, Seoul, 2014

satisfying assignments of a conjunctive formula, or the number of homomorphisms between relational structures.

As is easily seen, the unrestricted CSP is NP-complete, because it includes such well known NP-complete problems as Satisfiability and Graph Coloring. Therefore the CSP research focuses on restricted cases of this problem. The most fruitful way to restrict the general CSP is to fix the structure $\mathcal{H}$; then the problem reduces to the following question: given a relational structure $\mathcal{G}$, decide whether or not there exists a homomorphism from $\mathcal{G}$ to $\mathcal{H}$. The 'logic' form of the problem can be restricted in a similar way by fixing set D and set of predicates Γ that are allowed in the problem. The counting version of the CSP can also be restricted using the same approach. A CSP restricted in one of these ways is often referred to as *non-uniform* CSP and denoted by $\mathrm{CSP}(\mathcal{H})$ or $\mathrm{CSP}(\Gamma)$, respectively. For the counting version of the problem we use $\#\mathrm{CSP}(\mathcal{H})$ and $\#\mathrm{CSP}(\Gamma)$. Non-uniform CSPs allow one to express a wide range of combinatorial problems such as graph coloring, satisfiability, independent sets, etc.

The principal research goals in the study of non-uniform CSPs, decision or counting, are (1) to determine its precise complexity by identifying the complexity classes it is complete in, and (2) to design efficient solution algorithms whenever such algorithms exist. Decision non-uniform CSPs have received much attention during the last couple of decades, see, e.g., [2, 6, 7, 15, 43, 45, 56]. One of the most remarkable phenomena emerging from these studies and directly related to goal (1) is that every non-uniform CSP of known complexity is either polynomial time solvable or NP-complete. Such a property, often referred to as complexity dichotomy, is, however, nontrivial, since by [48] there is an infinite hierarchy of complexity classes between P and NP (assuming P$\neq$NP). The dichotomy phenomenon triggered a flurry of similar results, of which the counting CSP is an important part.

Another origin of the counting CSP can be traced back to statistical physics. Several formalisms such as the Ising and Potts models [44, 53] have been used to study macroscopic properties of ensembles of particles from their local interactions. Ising, Potts, and other models involve computing partition functions corresponding to a collection of particles. These models result in a generalized version of the counting CSP known as the weighted counting CSP, in which every homomorphism (satisfying assignment) is equipped with a real or complex weight, and the goal is to find the sum — the partition function — of the weights of all homomorphisms (or satisfying assignments). Although the goals in physics studying the models are different from the algorithmic and complexity questions we are concerned with, this link has served as a source of motivation for the counting CSP research.

2. Basic facts

2.1. Counting and approximation complexity. Counting problems including #CSPs as well as weighted #CSPs form a subclass of problems of computing functions. Recall that in such a problem the goal is, given a string $w \in \Sigma^*$, Σ is a finite alphabet and Σ^* is the set of finite strings over Σ, to find the value $f(w)$, where $f : \Sigma^* \to \mathbb{S}$ is a function to some set $\mathbb{S}$. In this paper $\mathbb{S}$ is always some numerical set, $\mathbb{N}$, the set of natural numbers, for counting CSPs, or $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, the sets of integers, rational numbers, real, and complex numbers, as well as, $\mathbb{Q}^+$, $\mathbb{R}^+$, the sets of positive rational and real numbers, respectively, for weighted #CSPs. Note that for algorithmic purposes elements of $\mathbb{S}$ must be efficiently representable. Although some approaches to complexity of arbitrary real and complex functions exist, see,

e.g. [47], we will avoid complications related to such methods by assuming that all numbers we use are algebraic and have an efficient representation.

FP denotes the class of functions f for which there exists a polynomial time deterministic Turing machine that computes f. Generally, the class of counting problems, #P, is defined as a counting analog of the class NP. A function $f\colon \Sigma^* \to \mathbb{N}$ belongs to the class #P if there exists a polynomial time non-deterministic Turing machine M such that for any instance $w \in \Sigma^*$ the number of accepting paths of $M(w)$ equals $f(w)$.

Parsimonious reduction between counting problems was introduced by Valiant in [62]. Problem (function) f is *parsimoniously reducible* to problem g if there is a polynomial time algorithm that transforms every instance w of f to an instance w' of g such that the size of w' is bounded by a polynomial in the size of w and $f(w) = g(w')$. Parsimonious reductions are often too restrictive to obtain meaningful results. In this paper we mostly use more general *Turing reductions*. Problem f is *Turing reducible* to problem g if there is a polynomial time algorithm computing f that satisfies the following condition: for every instance w of f it makes polynomially many queries of the form $g(w')$, where the size of w' is also bounded by a polynomial in the size of w.

#P-completeness with respect to parsimonious reductions is defined in a natural way: A counting problem f is *#P-complete* if it belongs to #P and any problem from #P is parsimoniously reducible to f. For Turing reductions the situation is more complicated. Technically, the class #P is not closed under Turing reductions, therefore, it is more correct to speak about $\mathrm{P}^{\#\mathrm{P}}$-completeness, that is, completeness in the class of problems solvable in polynomial time provided an access to an oracle from #P. This class contains very difficult problems, for example, Toda [59] showed that the *polynomial time hierarchy* is contained in $\mathrm{P}^{\#\mathrm{P}}$. To avoid unnecessary complications the hardness results in this paper will be stated in terms of #P-hardness rather than completeness. A problem f is *#P-hard* if every problem from #P is Turing reducible to f.

Ladner [48] showed that if P≠NP then there are infinitely many different complexity classes between P and NP. Ladner's result can be extended to counting complexity classes. Many results cited here claim that a counting CSP must be either in FP or #P-hard. This, however, is a nontrivial property that cannot be taken for granted.

2.2. Constraint satisfaction problem. Let D be a set and $n \in \mathbb{N}$. The set of all *n-tuples* over D is denoted by D^n; we will denote tuples in boldface, e.g., $\mathbf{a}$, and its ith entry will be denoted by $\mathbf{a}[i]$; n is also referred to as the *arity* of $\mathbf{a}$. An *n-ary* relation R is any subset of D^n; again n is called the arity of R, denoted $\mathrm{ar}(R)$. We do not distinguish between a relation R and the corresponding predicate, the function $R : D^n \to \{0, 1\}$ with $R(\mathbf{a}) = 1$ if and only if $\mathbf{a} \in R$. A *constraint language* is any set of relations over some set.

We first introduce the 'logic' version of the counting CSP. For $\mathbf{a} \in D^n$ and $\sigma\colon D \to B$ by $\sigma(\mathbf{a})$ we denote the tuple $(\sigma(\mathbf{a}[1]), \ldots, \sigma(\mathbf{a}[n]))$. For $n \in \mathbb{N}$ the set $\{1, \ldots, n\}$ is denoted by $[n]$.

Definition 2.1. Let D be a finite set called a *domain* and Γ a constraint language over D. An instance I of the *counting constraint satisfaction problem associated with* Γ ($\#\mathrm{CSP}(\Gamma)$) consists of a finite set of variables V and a conjunctive formula

$$R_1(\mathbf{x}_1) \wedge \ldots \wedge R_m(\mathbf{x}_m),$$

where $R_1, \ldots, R_m \in \Gamma$ and $\mathbf{x}_i$ is a tuple of variables from V of arity $\mathrm{ar}(R_i)$. A *solution* of I is a mapping $\sigma : V \to D$ such that $R_i(\sigma(\mathbf{x}_i)) = 1$ for every $i \in [m]$. The objective

in $\#\mathrm{CSP}(\Gamma)$ is to find the number of solutions of a given instance I. By the size of I we understand the length of any reasonable encoding of I.

Feder and Vardi [34] observed that CSP can be equivalently expressed through homomorphisms of relational structures.

A *vocabulary* is a finite set of relational symbols $R_1, \dots, R_n$ each of which has a fixed arity $\mathrm{ar}(R_i)$. A *relational structure* over vocabulary $R_1, \dots, R_n$ is a tuple $\mathcal{H} = (H; R_1^{\mathcal{H}}, \dots, R_n^{\mathcal{H}})$ such that H is a non-empty set, called the *universe* of $\mathcal{H}$, and each $R_i^{\mathcal{H}}$ is a relation on H having the same arity as the symbol R_i. Let $\mathcal{G}, \mathcal{H}$ be relational structures over the same vocabulary $R_1, \dots, R_n$. A *homomorphism* from $\mathcal{G}$ to $\mathcal{H}$ is a mapping $\sigma\colon G \to H$ from the universe G of $\mathcal{G}$ (the *instance*) to the universe H of $\mathcal{H}$ (the *template*) such that, for every relation $R^{\mathcal{G}}$ of $\mathcal{G}$ and every tuple $\mathbf{a} \in R^{\mathcal{G}}$, it holds that $\sigma(\mathbf{a}) \in R^{\mathcal{H}}$.

Definition 2.2. Let $\mathcal{H}$ be a relational structure. In the *counting constraint satisfaction problem associated with $\mathcal{H}$ ($\#\mathrm{CSP}(\mathcal{H})$)*, the objective is, given a structure $\mathcal{G}$ with the same vocabulary as $\mathcal{H}$, to compute the number of homomorphisms from $\mathcal{G}$ to $\mathcal{H}$. By the size of $\mathcal{G}$ we understand the length of any reasonable encoding of $\mathcal{G}$.

Example 2.3 ($\#H$-Coloring, [30, 40, 49]). A graph $H = (V, E)$ is a structure with a vocabulary consisting of one binary symbol E. Then $\#\mathrm{CSP}(H)$ is widely known as the $\#H$-Coloring Problem, in which the objective is to compute the number of homomorphisms from a given graph to H.

The $\#H$-Coloring Problem can also be represented as $\#\mathrm{CSP}(\Gamma_H)$, where $\Gamma_H = \{E\}$. Indeed, every instance, i.e., a graph $G = (W, F)$ can be converted into a formula $\bigwedge_{uv \in F} E(u, v)$.

Example 2.4 (#3-SAT, [22, 23, 61, 62]). An instance of the #3-SAT problem is specified by giving a propositional logic formula in CNF each clause of which contains 3 literals, and asking how many assignments satisfy it. Therefore, #3-SAT is equivalent to $\#\mathrm{CSP}(\Gamma_{3\text{-SAT}})$, where $\Gamma_{3\text{-SAT}} = \{R_{ijk}\colon i, j, k \in \{0, 1\}\}$ and the $R_{ijk} = \{0, 1\}^3 - \{(i, j, k)\}$ are the ternary relations representable by 3-clauses.

Alternatively, #3-SAT can be represented as $\#\mathrm{CSP}(\mathcal{S}_3)$, where $\mathcal{S}_3$ is the 2-element relational structure with universe $\{0, 1\}$ and vocabulary $\{R_{ijk}\colon i, j, k \in \{0, 1\}\}$.

Example 2.5 (Systems of linear equations, [22, 23]). Let F be a finite field and let $\#3\mathrm{LIN}(F)$ be the problem of finding the number of solutions to a system of linear equations over F, each of which contains at most 3 variables. Similar to Example 2.4 this problem can be viewed as $\#\mathrm{CSP}(\Gamma_{3\mathrm{LIN}(F)})$ where $\Gamma_{3\mathrm{LIN}(F)}$ contains all relations representable by a linear equation with at most 3 variables.

For the homomorphism version of this problem, it is not hard to see that $\#3\mathrm{LIN}(F)$ is equivalent to $\#\mathrm{CSP}(\mathcal{L}_3)$, where $\mathcal{L}_3$ is the relational structure with the universe F each of whose relations is given by a linear equation with at most 3 variables.

The more general problem $\#\mathrm{LIN}(F)$, in which linear equations are not limited to have 3 variables, can also be expressed as $\#\mathrm{CSP}(\Gamma_{\mathrm{LIN}(F)})$ for an appropriate constraint language. However, $\Gamma_{\mathrm{LIN}(F)}$ is necessarily infinite.

In general, a relation that can be represented as the set of solutions of a system of linear equations over a field will be called *affine*.

Example 2.6 (Independent sets). In the #Independent Set problem (#IS) the objective is, given a graph G, to find the number of independent sets in G. As is easily seen, #IS is equivalent to $\#\mathrm{CSP}(\mathcal{I})$, where $\mathcal{I} = (\{0, 1\}, E_1)$ is a graph with $E_1 = \{(0, 0), (0, 1), (1, 0)\}$.

Since #IS is a special case of the $\#H$-COLORING problem, its representation as a CSP in the 'logic' form is the same as in Example 2.3.

We now turn to weighted #CSPs. Recall that $\mathbb{S}$ is one of the sets $\mathbb{Z}, \mathbb{N}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, $\mathbb{Q}^+, \mathbb{R}^+$, although the following definition works for any semiring. For a finite set D, the *domain*, we let

$$\mathcal{F}_k(D, \mathbb{S}) = \{f \colon D^k \to \mathbb{S}\}, \qquad \mathcal{F}(D, \mathbb{S}) = \bigcup_{k \geq 1} \mathcal{F}_k(D, \mathbb{S})$$

denote the set of functions from D to $\mathbb{S}$. If $f \in \mathcal{F}_k(D, \mathbb{S})$, then k is called the arity of f, denoted $\mathrm{ar}(f)$.

Definition 2.7. Let $\Gamma \subseteq \mathcal{F}(D, \mathbb{S})$ be a finite set. The weighted #CSP associated with Γ, denoted $\#\mathrm{CSP}(\Gamma)$, is the following problem. An instance I of $\#\mathrm{CSP}(\Gamma)$ consists of a finite set of variables V and a product

$$f_1(\mathbf{x}_1) \cdot \ldots \cdot f_m(\mathbf{x}_m),$$

where $f_1, \ldots, f_m \in \Gamma$ and $\mathbf{x}_i$ is a tuple of variables from V of arity $\mathrm{ar}(f_i)$. A *configuration* σ for the instance I is a function $\sigma : V \to D$. The *weight* of a configuration σ is given by

$$w(\sigma) = \prod_{i=1}^{m} f_i(\sigma(\mathbf{x}_i)).$$

The objective in $\#\mathrm{CSP}(\Gamma)$ is to find the *partition function* of I

$$Z_\Gamma(I) = \sum_{\sigma : V \to D} w(\sigma).$$

An important special case of weighted #CSP is when Γ contains only one function f, and this function is binary. Let $D = \{d_1, \ldots, d_\ell\}$. Then f can be described by an $\ell \times \ell$-matrix M_f where $M_f[i, j] = f(d_i, d_j)$. The problem $\#\mathrm{CSP}(\{f\})$ (or $\#\mathrm{CSP}(f)$ for simplicity) is often referred to as $\mathrm{EVAL}(M_f)$. In other words, for an $\ell \times \ell$-matrix M the problem $\mathrm{EVAL}(M)$ is the problem of computing, given a multi-digraph $G = (V, E)$, the sum

$$Z_M(G) = \sum_{\sigma : V \to [\ell]} \prod_{uv \in E} M[\sigma(u), \sigma(v)].$$

Vertex weights add flexibility to the problem $\mathrm{EVAL}(M)$ and often occur in proofs as an auxiliary tool. Technically, $\mathrm{EVAL}(M)$ with vertex weights, denoted $\mathrm{EVAL}(M, N)$, is the problem $\#\mathrm{CSP}(f, g)$, where f is the binary function corresponding to M, and g is a unary function defining the weights. It is also standard to model g by a diagonal matrix N such that $N[i, i] = g(d_i)$. The partition function then takes the following shape

$$Z_{M,N}(G) = \sum_{\sigma : V \to [\ell]} \prod_{uv \in E} M[\sigma(u), \sigma(v)] \cdot \prod_{u \in V} N[\sigma(u), \sigma(u)].$$

The following two examples originate from certain "spin-glass" models of statistical physics.

Example 2.8 (The Ising model, [44]). This framework was introduced by Ising [44] to describe the phase transition phenomenon in ferromagnets. For a given graph $G = (V, E)$, the model assigns a *spin* $\sigma(v)$ to every vertex $v \in V$, which may be -1 or $+1$. Each edge contributes some amount of energy into the system, this amount is the same for every edge and equals $-J\sigma(u)\sigma(v)$ for some constant J; also every vertex contributes the same amount of energy N that is determined by the external magnetic field. Then the total energy of the system (its Hamiltonian) equals $H(\sigma) = -J\sum_{uv\in E}\sigma(u)\sigma(v) - N\sum_{v\in V}\sigma_v$. Let T denote the temperature of the system, and let k denote Bolzmann's constant. Then setting $\beta = (kT)^{-1}$, the partition function of the system is given by

$$Z_{\text{Ising}}(G,T) = \sum_{\sigma:V\to\{-1,+1\}} e^{-\beta H(\sigma)} = \sum_{\sigma:V\to\{-1,+1\}} \prod_{uv\in E} e^{\beta J\sigma(u)\sigma(v)} \prod_{v\in V} e^{\beta N\sigma(v)}.$$

Defining function $f_{\text{Ising}}(x,y)$ by $f_{\text{Ising}}(-1,-1) = f_{\text{Ising}}(+1,+1) = e^{2\beta J}$ and $f_{\text{Ising}}(-1,+1) = f_{\text{Ising}}(+1,-1) = 1$, function $g_{\text{Ising}}(x) = e^{\beta N x}$, and $\Gamma_{\text{Ising}} = \{f_{\text{Ising}}, g_{\text{Ising}}\}$ we have $Z_{\text{Ising}}(G,T) = e^{\beta J|E|}Z_{\Gamma_{\text{Ising}}}(G)$.

Alternatively, computing $Z_{\text{Ising}}(G,T)$ is equivalent to $\text{EVAL}(M_{\text{Ising}}, N_{\text{Ising}})$ where

$$M_{\text{Ising}} = \begin{pmatrix} e^{2\beta J} & 1 \\ 1 & e^{2\beta J} \end{pmatrix}, \qquad N_{\text{Ising}} = \begin{pmatrix} e^{-\beta N} & 0 \\ 0 & e^{\beta N} \end{pmatrix},$$

Example 2.9 (The Potts model, [53]). The Potts model was introduced in [53] as a generalization of the Ising model. For a natural number q, in the *q-state Potts model* on a graph $G = (V, E)$ every configuration $\sigma : V \to [q]$ has the Hamiltonian given by $H(\sigma) = -J\sum_{uv\in E}\delta_{\sigma(u)\sigma(v)}$, where δ is the Kronecker delta. Setting $x = e^{\beta J} - 1$ the corresponding partition function Z_{Potts} satisfies the equality

$$Z_{\text{Potts}}(G;q,x) = \sum_{\sigma:V\to[q]} e^{-\beta H(\sigma)} = \sum_{\sigma:V\to[q]} \prod_{uv\in E} (1 + x\cdot\delta_{\sigma(u),\sigma(v)}) = Z_{\Gamma_{\text{Potts}}}(G),$$

where $\Gamma_{\text{Potts}} = \{f_{\text{Potts}}\}$ and $f_{\text{Potts}} : [q]\times[q] \to \mathbb{R}$ is given by $f_{\text{Potts}}(x,y) = e^{\beta J}$ if $x = y$ and $f_{\text{Potts}}(x,y) = 1$ otherwise.

Example 2.10 (Even Subgraphs, [36]). Let us consider the problem $\text{EVAL}(M_{\text{Even}})$ for the following 2×2-matrix, whose rows and columns are indexed with $0, 1$

$$M_{\text{Even}} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

For a graph $G = (V, E)$ the value $\frac{1}{2}Z_{M_{\text{Even}}}(G) + 2^{|V|-1}$ is the number of induced subgraphs of G with an even number of edges ([36]). To see this, observe that for every configuration $\sigma : V \to \{0, 1\}$ the term $\prod_{uv\in E} M_{\text{Even}}[\sigma(u), \sigma(v)]$ is 1 if the subgraph induced by $\{v\colon \sigma(v) = 1\}$ has an even number of edges and -1 otherwise. Therefore up to a simple transformation, $Z_{M_{\text{Even}}}$ counts induced subgraphs with an even number of edges.

3. Ranks, linear equations, and groups

Many counting problems studied since the foundational results by Valiant [61, 62] are problems of the form $\#\text{CSP}(\Gamma)$, see, e.g. [26, 52, 54]. Problems related to the Potts and

Ising models have been studied in statistical physics and elsewhere for nearly a century [44, 46, 53]. However, a systematic study of the complexity of (weighted) #CSP has started only in the late 90s.

In this section we present the three main algorithmic ideas used to solve #CSPs.

3.1. Linear equations. In Example 2.5 for a finite field F we introduced the problems $\#3\mathrm{LIN}(F)$ and $\#\mathrm{LIN}(F)$. Observe that the number of solutions of a system of linear equations over F can be easily found. Indeed, if the dimensionality of the solution space of a given system equals k, the system has $|F|^k$ solutions. More surprising is that in some cases this is the only reason a #CSP can be solved in polynomial time.

By a *Boolean* CSP (#CSP, weighted #CSP) we mean a problem whose domain D contains only two elements. We will assume $D = \{0, 1\}$. Creignou and Herrmann [22, 23] showed that among Boolean #CSPs the only polynomial time solvable case is $\mathrm{LIN}(F)$, where F is the 2-element field.

Theorem 3.1 ([22, 23]). *Let* Γ *be a set of relations on* $\{0, 1\}$. *Then* $\#\mathrm{CSP}(\Gamma) \in \mathrm{FP}$ *if and only if every relation from* Γ *is* affine. *Otherwise it is #P-hard.*

3.2. Graphs and ranks. Consider the problem $\mathrm{EVAL}(M)$ where M is an $\ell \times \ell$-matrix with entries from $\mathbb{S}$ (recall that $\mathbb{S}$ denotes one of $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{Q}^+, \mathbb{R}, \mathbb{R}^+, \mathbb{C}$), and let $\mathsf{rank}(M)$ denote the rank of M. As is mentioned before, $\mathrm{EVAL}(M)$ is the same as the problem $\#\mathrm{CSP}(f)$ for an appropriate binary function f.

Lemma 3.2. *If* $\mathsf{rank}(M) \leq 1$ *then* $\mathrm{EVAL}(M)$ *is solvable in polynomial time.*

Proof. An instance of $\mathrm{EVAL}(M)$ is a multi-digraph $G = (V, E)$, and

$$Z_M(G) = \sum_{\sigma: V \to D} \prod_{uv \in E} M[\sigma(u), \sigma(v)].$$

Let $\mathsf{indeg}(u), \mathsf{outdeg}(u)$ denote the indegree and outdegree of a vertex $v \in V$. Let $|D| = k$. As $\mathsf{rank}(M) \leq 1$ there are numbers $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k \in \mathbb{S}$ such that for $1 \leq i, j \leq k$ we have:

$$M[i, j] = \alpha_i \cdot \beta_j$$

(β_j can be chosen to be the $M[1, j]$ and $\alpha_i = M[i, 1]/M[1, 1]$). Then for $\sigma : V \to D$

$$\omega(\sigma) = \prod_{uv \in E} M[\sigma(u), \sigma(v)] = \prod_{uv \in E} \alpha_{\sigma(u)} \beta_{\sigma(v)} = \prod_{v \in V} \alpha_{\sigma(v)}^{\mathsf{outdeg}(v)} \beta_{\sigma(v)}^{\mathsf{indeg}(v)}.$$

Thus

$$Z_M(G) = \sum_{\sigma: V \to D} \omega(\sigma) = \sum_{\sigma} \prod_{v \in V} \alpha_{\sigma(v)}^{\mathsf{outdeg}(v)} \beta_{\sigma(v)}^{\mathsf{indeg}(v)} = \prod_{v \in V} \sum_{i=1}^{k} \alpha_i^{\mathsf{outdeg}(v)} \beta_i^{\mathsf{indeg}(v)}.$$

The last term can easily be evaluated in polynomial time. □

If all entries of M are natural numbers, M can be viewed as the adjacency matrix of a multi-digraph $H_M = (V, E)$. In this case $\mathrm{EVAL}(M)$ is also the problem of counting the number of homomorphisms of a given multi-digraph $G = (W, F)$ to H_M. By a homomorphism of multi-digraphs here we understand a mapping $\sigma : F \to E$ such that for any arcs

$uv, wt \in F$ with $\sigma(uv) = u'v', \sigma(wt) = w't'$, if $u = w$ then $u' = w'$, and if $v = t$ then $v' = t'$.

Another way to interpret the problem $\mathrm{EVAL}(M)$ for matrices with entries from $\mathbb{N}$ is through equivalence relations. Let η, θ be equivalence relations on a set D and let $A_1, \dots, A_k$ and $B_1, \dots, B_\ell$ be the equivalence classes of η and θ, respectively. The matrix $M_{\eta,\theta}$ is a $k \times \ell$-matrix such that $M_{\eta,\theta}[i,j] = |A_i \cap B_j|$.

The connection between $\#\mathrm{CSP}(\eta, \theta)$, EVAL, and weighted $\#$CSP was established in [14]. The *bipartization* of a $k \times \ell$-matrix M is the $(k+\ell) \times (k+\ell)$-matrix

$$\mathsf{bip}(M) = \begin{pmatrix} 0 & M \\ 0 & 0 \end{pmatrix}.$$

Matrix $\mathsf{bip}(M)$ is always the adjacency matrix of a bipartite (multi)-digraph with parts A and B of size k and ℓ, respectively, and such that every edge is directed from A to B. The problems $\#\mathrm{CSP}(\eta, \theta)$ and $\mathrm{EVAL}(\mathsf{bip}(M_{\eta,\theta}))$ are Turing reducible to each other. If $M_{\eta,\theta}$ is a square matrix then $\#\mathrm{CSP}(\eta, \theta)$ is reducible to $\mathrm{EVAL}(M_{\eta,\theta})$; and for arbitrary η, θ, the problem $\mathrm{EVAL}(M_{\eta,\theta} \cdot M_{\eta,\theta}^{\top})$ is reducible to $\#\mathrm{CSP}(\eta, \theta)$.

The rank 1 condition from Lemma 3.2 can be generalized further. Let M be a $k \times \ell$-matrix with entries from $\mathbb{S}$. A *submatrix* of M is a matrix obtained from M by deleting some rows and columns. For non-empty sets $I = \{i_1, \dots, i_p\} \subseteq [k]$ and $J = \{j_1, \dots, j_q\} \subseteq [\ell]$, by M_{IJ} we denote the $p \times q$-submatrix with $M_{IJ}[r,s] = M[i_r, j_s]$. A submatrix M' is *proper* if $M' \neq M$. Let $\overline{I} = [k] - I$ and $\overline{J} = [\ell] - J$.

Definition 3.3. A *decomposition* of M consists of two proper submatrices M_{IJ} and $M_{\overline{I}\,\overline{J}}$ such that $M[i,j] = 0$ for all $(i,j) \in (I \times \overline{J}) \cup (\overline{I} \times J)$. Matrix M is *indecomposable* if it has no decomposition. A *block* of M is an indecomposable submatrix M_{IJ} with at least one non-zero entry and such that $M_{IJ}, M_{\overline{I}\,\overline{J}}$ is a decomposition of M.

The following theorem by Bulatov and Grohe [14] characterizes matrices that give rise to counting CSPs solvable in polynomial time. If every block of a matrix M has rank 1, M is said to be *block rank-one*.

Theorem 3.4 ([14]).

(1) *Let η, θ be equivalence relations on D and $M = M_{\eta,\theta}$ the corresponding matrix. Then the problem $\#\mathrm{CSP}(\eta, \theta)$ is solvable in polynomial time if and only if M block rank-one. Otherwise this problem is #P-hard.*

(2) *Let $f : D^2 \to \mathbb{R}^+$ be a commutative function and $M = M_f$ (or $M = M_H$ for a multi-graph H). Then $\#\mathrm{CSP}(f)$ (resp, or $\mathrm{EVAL}(M_H)$) is solvable in polynomial time if and only if M is block rank-one. Otherwise this problem is #P-hard.*

The block rank-one condition can be generalized to tensors of more than two dimensions. Let M be an r-dimensional tensor over $\mathbb{S}$ indexed by elements of $[k_1] \times \ldots \times [k_r]$. It is said to have *rank 1* if there are numbers γ_{ij}, $i \in [r]$, $j \in [k_i]$ such that $M[i_1, \dots, i_r] = \gamma_{1i_1} \cdot \ldots \cdot \gamma_{ri_r}$. Decompositions and blocks of M are defined similar to Definition 3.3. Finally, M has the block rank-one condition if every its block has rank at most one.

The block rank-one condition and Theorem 3.4 provide some insights into more complex cases.

Undirected graphs. Dyer and Greenhill [30] proved a dichotomy theorem for #CSP over undirected graphs.

Theorem 3.5 ([30]). *Let H be a graph, possibly with loops. Then $\#\mathrm{CSP}(H) \in \mathrm{FP}$ if and only if every connected component of H is either a single vertex, or a complete graph with all loops present, or a complete bipartite graph without loops. Otherwise $\#\mathrm{CSP}(H)$ is #P-hard.*

As is easily seen, the adjacency matrices of graphs from Theorem 3.5 are the symmetric 01-matrices M satisfying the block rank-one condition.

The following two results from [29] and [31] make use of Theorem 3.4, although the problems they study do not involve matrices explicitly. The reductions to binary functions constructed in these papers are somewhat ad hoc; however, we shall see in the next section how such reductions can be introduced and studied more systematically.

Directed acyclic graphs. Dyer et al. in [29] considered the complexity of $\#\mathrm{CSP}(H)$, where $H = (V, E)$ is a *directed acyclic digraph* (DAG). A DAG H is said to be *layered* if there is a partition $V_0, V_1, \ldots, V_\ell$ of V such that $uv \in E$ if and only if $u \in V_i$ and $v \in V_{i+1}$ for some i, or, equivalently, for any two nodes u, v all directed paths from u to v have the same length. Let $H^{[i,j]}$ $(i < j)$ denote the subgraph of H induced by $V_i \cup \ldots \cup V_j$. For $u \in V_i$ and $v \in V_j$ $(i < j)$ by H_{uv} we denote the subgraph of H induced by those connected components of $H^{[i+1,j-1]}$ to which both u and v are connected. For an (acyclic) digraph G let A_G denote the $|V| \times |V|$-matrix such that $A_G[u, v]$ equals the number of homomorphisms from G to H_{uv} if $u \in V_i$, $v \in V_j$ with $i < j$, and $A_G[u, v] = 0$ otherwise.

Theorem 3.6 ([29]). *Let H be a DAG. Then $\#\mathrm{CSP}(H) \in \mathrm{FP}$ if and only if H is layered and for any digraph G the matrix A_G is block rank-one. Otherwise the problem is #P-hard.*

Clearly, the condition given in Theorem 3.6 cannot be straightforwardly decided. However, using Lovasz's result [51] that digraphs H_1, H_2 are isomorphic if and only if any digraph G has the same number of homomorphisms to H_1 and to H_2, the condition can be much simplified. It is, in fact, equivalent to the condition that for any u, u', v, v' such that u, u' and v, v' are on the same level, the graphs $H_{uv}H_{u'v'}$ and $H_{uv'}H_{u'v}$ are isomorphic. Here HG, for $H = (V, E)$, $G = (W, F)$, denotes the product of graphs, that is, the graph with vertex set $V \times W$ and such that $(u, u')(v, v')$ is an edge if and only if $uv \in E$ and $u'v' \in F$.

Hypergraphs. To explain the result of [31] we need several definitions. Recall that a *hypergraph* is a set of vertices V and a collection E of *hyperedges*, subsets of vertices from V. A hypergraph $H = (V, E)$ is said to be *r-uniform* if every hyperedge from E contains exactly r elements. A homomorphism of a hypergraph $G = (W, F)$ to H is a mapping $\sigma : W \to V$ such that the image $\sigma(A)$ of any hyperedge $A \in F$ is a hyperedge from E. The focus of [31] is on the complexity of $\#\mathrm{CSP}(H)$, the problem of counting homomorphisms to a r-uniform hypergraph H. It also considers an extension of this problem where every hyperedge from E is associated with a positive weight. In other words the problem considered in [31] can be stated as $\#\mathrm{CSP}(f)$, where f is an r-ary function $V^r \to \mathbb{R}^+$, which is *totally symmetric*, that is, $f(x_1, \ldots, x_r) = f(x_{\pi(1)}, \ldots, x_{\pi(r)})$ for any permutation π of $[r]$, and $f(a_1, \ldots, a_r) = 0$ whenever $a_1, \ldots, a_r \in V$ are not all different.

Consider D as a graph with vertex set D and edge set E' such that $ab \in E'$ if and only if there are $c_3, \dots, c_r \in D$ with $f(a, b, c_3, \dots, c_r) \neq 0$. We denote this graph by H_f. Since f is symmetric H_f is undirected. Let $D_1, \dots, D_m$ be the connected components of H_f. Function f will be be said to have an *affine decomposition* if it satisfies the following conditions:

AD1 Every D_i can be represented as a Cartesian product $[q_i] \times A_i$.

AD2 For every $i \in [m]$ and every $j \in [q_i]$ there is $\gamma_{ij} \in \mathbb{R}^+$ such that whenever $f(d_1, \dots, d_r) \neq 0$ and all $d_j = (t_j, a_j) \in D_i$ for some $i \in [m]$, $f(d_1, \dots, d_r) = \gamma_{it_1} \dots \gamma_{it_r}$.

AD3 For every $i \in [m]$ there exists an Abelian group with universe A_i and the operation of addition $+_i$, and $c_i \in A_i$ such that for any $d_1, \dots, d_r \in D_i$, $d_j = (t_j, a_j)$, the function $f(d_1, \dots, d_r) \neq 0$ if and only if $a_1 +_i \dots +_i a_r = c_i$.

AD4 In all other cases $f(d_1, \dots, d_r) = 0$.

Theorem 3.7 ([31]). *Let H be an r-uniform hypergraph with weights, and let $f : D^r \to \mathbb{R}^+$ be the corresponding totally symmetric function such that $f(d_1, \dots, d_r) = 0$ whenever not all the d_i are different. The problem $\#\mathrm{CSP}(f)$ is solvable in polynomial time if and only if f has an affine decomposition. Otherwise the problem is #P-hard.*

3.3. The group condition. If matrix M contains negative or complex values, the problem $\mathrm{EVAL}(M)$ can be solvable in polynomial time even though M does not satisfy the block rank-one condition, due to cancellations possible in the presence of negative and complex numbers. New interesting counting algorithms may appear.

Example 3.8. We consider the following simple binary function f on $\{0, 1\}$, see Example 2.10. The values of f are given by the matrix $M[x, y]$ whose rows and columns are indexed with $0, 1$, and $M[x, y]$ is the value of the function on x, y:

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Clearly, instances of the corresponding counting CSP are undirected multi-graphs

Since $f(x, y) = (-1)^{xy}$, given a graph $G = (V, E)$, $V = \{v_1, \dots, v_n\}$, we have

$$Z_M(G) = \sum_{\sigma: V \to \{0,1\}} \prod_{v_i v_j \in E} (-1)^{\sigma(v_i)\sigma(v_j)} = \sum_{\sigma: V \to \{0,1\}} (-1)^{g(\sigma(v_1), \dots, \sigma(v_n))} = S_0 - S_1,$$

where $g(x_1, \dots, x_n) = \sum_{v_i v_j \in E} x_i x_j$ is a quadratic polynomial over $GF(2)$, and S_a, $a = 0, 1$, is the number of solutions of the equation $g(x_1, \dots, x_n) = a$.

A method to compute the number of solutions of a quadratic equation over $GF(2)$ can be found in [50]. It is based on the following observation. Quadratic polynomials $f(x_1, \dots, x_n)$ and $g(x_1, \dots, x_n)$ are said to be *equivalent* if one can be obtained from the other by a linear substitution of indeterminates. In particular, equivalent polynomials have the same number of roots. Every (nontrivial) quadratic polynomial $f(x_1, \dots, x_n)$ is equivalent to a polynomial of the form $x_1 x_2 + g(x_3, \dots, x_n)$ where g is a quadratic polynomial. Then the number of roots of f equals the number of roots of g (setting $x_1 = 0$) plus the number of roots of $x_2 + g(x_3, \dots, x_n)$ (setting $x_1 = 1$), that is, 2^{n-2}.

The observations made in Example 3.8 lie in the core of all complexity results on weighted #CSP, along with the block rank-one condition. In [12] it is used to determine the complexity of the weighted #CSP over a 2-element domain as follows.

Theorem 3.9 ([12]). *Let $\Gamma \subseteq \mathcal{F}(\{0,1\},\mathbb{R})$. The problem $\#\mathrm{CSP}(\Gamma)$ is solvable in polynomial time if and only if one of the following two conditions hold:*

(1) *Every function $f \in \Gamma$ has the form $f(\mathbf{x}) = w(-1)^{s(\mathbf{x})}g(\mathbf{x})$, where $w \in \mathbb{R}^+$, s is a degree 2 polynomial, and g is an affine predicate (see Example 2.5); or*

(2) *Every function $f \in \Gamma$ can be represented as $f(x_1,\ldots,x_n) = h_1(x_1)\ldots h(x_n)\cdot g(x_1,\ldots,x_n)$, where each h_i is a unary function, and g is a polynomial which is a product of binary polynomials of the form $x_i + x_j$ and $x_i + x_j + 1$. (Note that in this case f satisfies the block rank-one condition.)*

Otherwise $\#\mathrm{CSP}(\Gamma)$ is #P-hard.

For domains with more than 2 elements several results [18, 36, 58] focus on binary functions and their matrices. In most cases these results involve a sophisticated chain of reductions in order to identify the tractable cases. We do not describe these reductions in full details, instead we highlight the most important ones and introduce the core tractable cases.

Goldberg et al. in [36] considered real symmetric matrices. The key type of matrices in this case is a generalization of the small matrix we considered in Example 3.8, *Hadamard* matrices. Recall that a square $n \times n$-matrix is called Hadamard if all its entries are from $\{1,-1\}$ and $H \cdot H^T$ is a diagonal matrix, that is, the rows of H are pairwise orthogonal. Hadamard matrices that give rise to #CSPs solvable in polynomial time can again be described through quadratic polynomials. Let H be an Hadamard $n \times n$-matrix such that $n = 2^k$, and let $\varrho^R : GF(2)^k \to [n]$ and $\varrho^C : GF(2)^k \to [n]$ be bijective mappings, called *index mappings*. For $\mathbf{x} = (x_1,\ldots,x_k), \mathbf{y} = (y_1,\ldots,y_k)$, a polynomial $h(\mathbf{x},\mathbf{y})$ over $GF(2)$ *represents* H with respect to ϱ^R and ϱ^C if for all $\mathbf{a},\mathbf{b} \in GF(2)^k$ it holds that $H[\varrho^R(\mathbf{a}),\varrho^C(\mathbf{b})] = (-1)^{h(\mathbf{a},\mathbf{b})}$.

Theorem 3.10 ([36]). *Let H be an Hadamard $n \times n$-matrix. The problem $\mathrm{EVAL}(H)$ is solvable in polynomial time if and only if $n = 2^k$ for some $k > 0$, and there are index mappings $\varrho^R, \varrho^C : GF(2)^k \to [n]$ and a quadratic polynomial $h(\mathbf{x},\mathbf{y})$, $\mathbf{x} = (x_1,\ldots,x_k), \mathbf{y} = (y_1,\ldots,y_n)$ such that h represents H with respect to ϱ^R, ϱ^C. Otherwise $\mathrm{EVAL}(H)$ is #P-hard.*

For a general real matrix M it is then proved that either $\mathrm{EVAL}(M)$ is #P-hard, or there is an Hadamard matrix H such that $\mathrm{EVAL}(M)$ is interreducible with $\mathrm{EVAL}(H,N)$ with additional vertex weights N.

The results of [36] have been further generalized by Thurley [58] to *Hermitian* matrices.

4. Algebraic approach and the hardness of #CSP

The other component of dichotomy results is proving the hardness of problems that cannot be solved efficiently. In this section we introduce the algebraic approach to the counting CSPs that turned out to be very useful for this purpose and eventually led to a complete complexity classification of unweighted #CSPs in [8], its further improvements and simplifications [32, 33], and guided the way to complexity results on the weighted #CSP.

4.1. Primitive positive definitions. Let Γ be a set of relations (predicates) over a finite set D. A relation R over D is said to be *primitive-positive (pp-) definable* in Γ if $R(\mathbf{x}) = \exists \mathbf{y}\, \Phi(\mathbf{x}, \mathbf{y})$, where Φ is a conjunction that involves predicates from Γ and equality relations. The formula above is then called a *pp-definition* of R in Γ. A constraint language Δ is pp-definable in Γ if so is every relation from Δ.

Example 4.1. Let $K_3 = (\{a, b, c\}, E)$ be a 3-element complete graph. Its edge relation is the binary disequality relation on $\{a, b, c\}$. Then the pp-formula

$$\begin{aligned} S(x, y, z) \quad = \quad & \exists t, u, v, w (E(t, x) \wedge E(t, y) \wedge E(t, z) \wedge E(u, v) \wedge E(v, w) \\ & \wedge E(w, u) \wedge E(u, x) \wedge E(v, y) \wedge E(w, z)) \end{aligned}$$

defines the relation S that consists of all triples containing exactly 2 different elements from $\{a, b, c\}$.

A link between pp-definitions and reducibility between decision CSPs was first observed in [45]. It was later extended to counting CSPs in [10].

Theorem 4.2 ([10]). *Let Γ and Δ be constraint languages over a finite set D and Δ finite. If Δ is pp-definable in Γ then $\#\mathrm{CSP}(\Delta)$ is Turing reducible to $\#\mathrm{CSP}(\Gamma)$.*

It also follows from the results of [13] that a similar approach works for weighted #CSPs. Let now $\Gamma \subseteq \mathcal{F}(D, \mathbb{S})$. Let also $V = \{x_1, \dots, x_n\}$ be a set of variables. An atomic formula has the form $f(x_{i_1}, \dots, x_{i_r})$, r is the arity of f. A *primitive-positive summation formula* (a *pps-formula*) is a summation of a product of atomic formulas. More precisely for $V' = \{x_1, \dots, x_{n+m}\}$ a pps-formula ψ over Γ is

$$\psi = \sum_{x_{n+1}, \dots, x_{n+m} \in D} \prod_{j=1}^{s} \varphi_j, \tag{4.1}$$

where all φ_j are atomic formulas over Γ in the variables V'. The formula ψ defines a function

$$\psi(\mathbf{x}) = \sum_{x_{n+1}, \dots, x_{n+m} \in D} \prod_{j=1}^{s} \varphi_j(x_{j1}, \dots, x_{jr_j}),$$

where $\mathbf{x} = (x_1, \dots, x_n)$. A function that can be defined via a pps-formula using functions from Γ is said to be *pps-definable* in Γ. A set $\Delta \subseteq \mathcal{F}(D, \mathbb{S})$ is pps-definable in Γ if so is every function from Δ. If in Equation (4.1) $m = 0$, we say that the function is *pp-definable* in Γ. The following theorem is implicit in [13].

Theorem 4.3 ([13]). *Let $\Gamma, \Delta \subseteq \mathcal{F}(D, \mathbb{S})$ be such that Δ is finite and Δ pps-definable in Γ. Then $\#\mathrm{CSP}(\Delta)$ is parsimoniously reducible to $\#\mathrm{CSP}(\Gamma)$.*

One interesting example of what definitions can achieve is defining unary functions. Let $\delta_d : D \to \mathbb{S}$, $d \in D$, denote the function given by $\delta_d(d) = 1$ and $\delta_d(x) = 0$ otherwise. The presence of such functions in Γ allows to 'pin' certain variables in a #CSP instance. Say, if $f_1(\mathbf{x}) \cdot \ldots \cdot f_\ell(\mathbf{x})$ is an instance of $\#\mathrm{CSP}(\Gamma)$, then the answer for the instance $\delta_d(x) \cdot f_1(\mathbf{x}) \cdot \ldots \cdot f_\ell(\mathbf{x})$ is $\sum_{\sigma : V \to D, \sigma(x) = d} w(\sigma)$. Bulatov and Dalmau [10] proved that adding δ_d to an unweighted constraint language does not change its complexity. The following statement generalizes this result to the weighted case.

Theorem 4.4. *Let $\Gamma \subseteq \mathcal{F}(D, \mathbb{R}^+)$ and $\Gamma_{\mathrm{id}} = \Gamma \cup \{\delta_d \colon d \in D\}$. Then $\#\mathrm{CSP}(\Gamma_{\mathrm{id}})$ is Turing reducible to $\#\mathrm{CSP}(\Gamma)$.*

4.2. Polymorphisms and invariants. Primitive positive definability can be concisely characterized using polymorphisms. An operation $\varphi : D^k \to D$ is said to be a *polymorphism* of a relation $R \subseteq D^n$ if for any $\mathbf{a}_1, \dots, \mathbf{a}_k \in R$ the tuple $f(\mathbf{a}_1, \dots, \mathbf{a}_k)$ also belongs to R, where $f(\mathbf{a}_1, \dots, \mathbf{a}_k)$ stands for

$$(f(\mathbf{a}_1[1], \dots, \mathbf{a}_k[1]), \dots, f(\mathbf{a}_1[n], \dots, \mathbf{a}_k[n])).$$

Operation φ is a polymorphism of a constraint language Γ if it is a polymorphism of every relation from Γ.

Example 4.5. Let R be an affine relation, that is, R is the solution space of a system of linear equations over a field F. Then the operation $\varphi(x, y, z) = x - y + z$ is a polymorphism of R. Indeed, let $A \cdot \mathbf{x} = \mathbf{b}$ be the system defining R, and $\mathbf{x}, \mathbf{y}, \mathbf{z} \in R$. Then

$$A \cdot \varphi(\mathbf{x}, \mathbf{y}, \mathbf{z}) = A \cdot (\mathbf{x} - \mathbf{y} + \mathbf{z}) = A \cdot \mathbf{x} - A \cdot \mathbf{y} + A \cdot \mathbf{z} = \mathbf{b}.$$

In fact, the converse can also be shown: if R is invariant under φ, where φ is defined in a certain finite field F then R is the solution space of some system of linear equations over F.

A link between polymorphisms and pp-definability of relations is given by *Galois connection*.

Theorem 4.6 (Galois connection, [5, 35]). *Let Γ be a constraint language on D, and let $R \subseteq D^n$ be a non-empty relation. Then R is preserved by all polymorphisms of Γ if and only if R is pp-definable in Γ.*

From the counting complexity point of view the most interesting type of polymorphisms is *Mal'tsev polymorphisms*, that is, ternary operations φ satisfying the equations $\varphi(x, y, y) = \varphi(y, y, x) = x$. The existence of a Mal'tsev polymorphism imposes strong structural conditions on a relation.

Let $R \subseteq D^n$ be a relation, $\mathbf{a} \in D^n$, and $I = \{i_1, \dots, i_k\} \subseteq [n]$. By $\mathrm{pr}_I \mathbf{a}$ we denote the tuple $(\mathbf{a}[i_1], \dots, \mathbf{a}[i_k])$ and by $\mathrm{pr}_I R$ the k-ary relation $\{\mathrm{pr}_I \mathbf{a} \colon \mathbf{a} \in R\}$. If $\mathbf{a} \in \mathrm{pr}_I R$ and $\mathbf{b} \in \mathrm{pr}_{[n]-I} R$, by $(\mathbf{a}, \mathbf{b})$ we denote the tuple $\mathbf{c} \in D^n$ such that $\mathbf{c}[i] = \mathbf{a}[i]$ if $i \in I$ and $\mathbf{c}[i] = \mathbf{b}[i]$ otherwise. We say that R is *rectangular*, if for any $I \subset [n]$ and any $\mathbf{a}, \mathbf{b} \in \mathrm{pr}_I R$, $\mathbf{c}, \mathbf{d} \in \mathrm{pr}_{[n]-I} R$, if the tuples $(\mathbf{a}, \mathbf{c}), (\mathbf{b}, \mathbf{c}), (\mathbf{b}, \mathbf{d})$ belong to R, the tuple $(\mathbf{a}, \mathbf{d})$ also belongs to R. A constraint language Γ is said to be *strongly rectangular* [33] if every relation pp-definable in Γ is rectangular.

Theorem 4.7 (folklore). *A constraint language is strongly rectangular if and only if it has a Mal'tsev polymorphism*[1].

The absence of a Mal'tsev polymorphism of a constraint language Γ implies the hardness of the counting CSP.

Theorem 4.8 ([10]). *If a constraint language Γ does not have a Mal'tsev polymorphism, then $\#\mathrm{CSP}(\Gamma)$ is #P-hard.*

[1]The most accessible source for a proof of this statement is probably [33]

4.3. Congruence singular and strongly balanced languages. The block rank-one condition, a necessary condition for the tractability of certain counting CSPs can also be generalized using pp-definitions. Here we give two such generalizations, from [8] and from [32, 33]. It is easy to see that these two conditions are equivalent, for a formal proof see [33]. The conditions we introduce are also equivalent to polynomial time solvability of #CSPs, as we show in the next section.

First we introduce the notion of congruence singularity [8]. Let Γ be a constraint language and R an n-ary relation pp-definable in Γ. As is easily seen, any $2n$-ary relation S such that $\mathrm{pr}_{[n]}S = \mathrm{pr}_{\{n+1,\dots,2n\}}S = R$ can be viewed as a binary relation on R. Such a relation S is called a *congruence* of R if it is pp-definable in Γ and is an equivalence relation on R. Let η, θ, κ be congruences of R such that $\kappa \subseteq \eta, \theta$. We associate with R and η, θ, κ a matrix $M_{R;\eta,\theta,\kappa}$ defined as follows. Let $A_1, \dots, A_k$ be the η-blocks, and let $B_1, \dots, B_\ell$ be the θ-blocks. Then $M_{R;\eta,\theta,\kappa}$ is a $k \times \ell$-matrix such that $M_{R;\eta,\theta,\kappa}[i, j]$ equals the number of κ-blocks inside $A_i \cap B_j$. A constraint language Γ is said to be *congruence singular* if for any R pp-definable in Γ, and any its congruences η, θ, κ, $\kappa \subseteq \eta, \theta$, the matrix $M_{R;\eta,\theta,\kappa}$ is block rank-one.

Next, we give a definition of balanced constraint languages as in [32, 33]. Let again R be an n-ary relation pp-definable in Γ, and let $k, \ell \le n$. Relation R can be naturally viewed as a ternary relation on $D^k \times D^\ell \times D^{n-k-\ell}$. Now, a ternary relation $S \subseteq A_1 \times A_2 \times A_3$ is said to be *balanced* if the matrix $M[x, y] = |\{z\colon (x, y, z) \in S\}|$ is a block rank-one matrix. A relation of arity $n \ge 3$ is balanced if every expression of it as a ternary relation on $D^k \times D^\ell \times D^{n-k-\ell}$ is balanced. We will say that Γ is *strongly balanced* if every relation pp-definable in Γ is balanced.

Example 4.9 (Directed acyclic graphs). Let H be a layered DAG, and let $V_0, \dots, V_\ell$ be its layers. We show that if H is congruence singular then the matrix A_G (see Section 3.2) is block rank-one for any digraph G.

Let $\Phi(G, H)$, $G = (W, F)$, $H = (V, E)$, and $W = \{w_1, \dots, w_n\}$, denote the set of homomorphisms from G to H viewed as a $|W|$-ary relation on V; each tuple represents a homomorphism. Then $\Phi(G, H)$ is pp-definable in H. Indeed,

$$\Phi(G, H)(w_1, \dots, w_n) = \bigwedge_{uv \in F} E(u, v).$$

If G is not layered then $\Phi(G, H)$ is empty; so assume G is layered. Let W_1, W_2 denote the set of vertices on the highest and lowest layers of G, respectively. Let η_1, η_2 be congruences of $\Phi(G, H)$ such that $(\sigma, \sigma') \in \eta_i$, $i = 1, 2$, iff $\sigma(v) = \sigma'(v)$ for all $v \in W_i$. By H_{v*}, $v \in V_i$, we denote the subgraph of $H^{[i+1,\ell]}$ induced by the connected components to which there is a directed path from v; similarly, H_{*w}, $w \in V_j$, denotes the subgraph of $H^{[0,j-1]}$, induced by the connected components from which there is a directed path to w; then, $H_{vw} = H_{v*} \cap H_{*w}$. As is easily seen the sets of the form $P_{u*} = \{\sigma \in \Phi(G, H)\colon \sigma(z) \in H_{u*}$ for $z \in W$ and $\sigma(z) = u$ for $z \in W_1\}$ are classes of η_1, the sets of the form $P_{*w} = \{\sigma \in \Phi(G, H)\colon \sigma(z) \in H_{*w}$ for $z \in W$ and $\sigma(z) = w$ for $z \in W_2\}$ are classes of η_2, and the sets of the form $P_{uw} = \{\sigma \in \Phi(G, H)\colon \sigma(z) \in H_{uw}$ for $z \in W, \sigma(z) = u$ for $z \in W_1$, and $\sigma(z) = w$ for $z \in W_2\}$ are classes of $\eta_1 \wedge \eta_2$ (although there are classes of those congruences not representable in the form P_{u*}, P_{*w}, P_{uw}).

Pick $u, u' \in V_i$, $v, v' \in V_j$, $1 \le i < j \le \ell$. We need to show that $H_{uv}H_{u'v'}$ and $H_{uv'}H_{u'v}$ are isomorphic. We use an observation made in [29] that $|\Phi(G, H_1H_2)| =$

$|\Phi(G,H_1)|\cdot|\Phi(G,H_2)|$. Since H is congruence singular, the matrix $M_{\Phi(G,H);\eta_1,\eta_2,=}$, where $=$ denotes the equality relation, is block rank-one. Hence

$$\begin{vmatrix} |\Phi(G,H_{uv})| & |\Phi(G,H_{uv'})| \\ |\Phi(G,H_{u'v})| & |\Phi(G,H_{u'v'})| \end{vmatrix} = 0,$$

or $\Phi(G,H_{uv}),\Phi(G,H_{u'v'})$ or $\Phi(G,H_{uv'}),\Phi(G,H_{u'v})$ are in different blocks of $M_{\Phi(G,H);\eta_1,\eta_2,=}$. In the latter case either $|\Phi(G,H_{u'v})| = |\Phi(G,H_{uv'})| = 0$ or $|\Phi(G,H_{uv})| = |\Phi(G,H_{u'v'})| = 0$. The result follows.

4.4. The complexity of #CSP and universal algebra. Recall that a (*universal*) *algebra* is an ordered pair $\mathbb{A} = (A,F)$ where A is a non-empty set and F is a family of finitary operations on A. Every constraint language on a set D can be associated with an algebra $\mathsf{Alg}(\Gamma) = (D,\mathsf{Pol}(\Gamma))$, where $\mathsf{Pol}(\Gamma)$ denotes the set of all polymorphisms of Γ.

For the algebraic terminology and basic properties see [37].

Definition 4.10.

(1) Let $\mathbb{A} = (A;F)$ be an algebra. The *k-th direct power* of $\mathbb{A}$ is the algebra $\mathbb{A}^k = (A^k;F)$ where we treat each operation $f \in F$ as acting on A^k component-wise.

(2) Let $\mathbb{A} = (A;F)$ be an algebra, and let B be a subset of A such that, for any (say, n-ary) $f \in F$, and for any $b_1,\dots,b_n \in B$, we have $f(b_1,\dots,b_n) \in B$. Then the algebra $\mathbb{B} = (B;F|_B)$, where $F|_B$ consists of restrictions of operations $f \in F$ onto B, is called a *subalgebra* of $\mathbb{A}$.

(3) Let $\mathbb{A}_1 = (A_1;F_1)$ and $\mathbb{A}_2 = (A_2;F_2)$ be such that $F_1 = \{f_i^1 \mid i \in I\}$, $F_2 = \{f_i^2 \mid i \in I\}$, and f_i^1, f_i^2 are of the same arity r_i, $i \in I$. A mapping $\varphi : A_1 \to A_2$ is called a *homomorphism* from $\mathbb{A}_1$ to $\mathbb{A}_2$ if $\varphi(f_i^1(a_1,\dots,a_{r_i})) = f_i^2(\varphi(a_1),\dots,\varphi(a_{r_i}))$ holds for all $i \in I$ and all $a_1,\dots,a_{r_i} \in A_1$. If the mapping φ is onto then $\mathbb{A}_2$ is said to be a *homomorphic image* of $\mathbb{A}_1$.

(4) A *congruence* of an algebra $\mathbb{A} = (A;F)$ is an equivalence relation on A invariant under all operations from F.

Let $\mathbb{A} = (A,F)$. It is called *#-tractable* if for any finite constraint language Γ on A such that $F \subseteq \mathsf{Pol}(\Gamma)$ the problem $\#\mathrm{CSP}(\Gamma)$ is solvable in polynomial time. Algebra $\mathbb{A}$ is *#P-hard* if there is a finite Γ with $F \subseteq \mathsf{Pol}(\Gamma)$ such that $\#\mathrm{CSP}(\Gamma)$ is #P-hard.

Theorem 4.11 ([10]). *Let $\mathbb{A} = (A;F)$ be a finite algebra. Then*

(1) *if $\mathbb{A}$ is #-tractable then so is every subalgebra, homomorphic image, and direct power of $\mathbb{A}$.*

(2) *if $\mathbb{A}$ has a #P-hard subalgebra, homomorphic image, or direct power, then $\mathbb{A}$ is #P-hard.*

For an algebra $\mathbb{A}$ the class of algebras that are homomorphic images of subalgebras of direct powers of $\mathbb{A}$ is called the *variety* generated by $\mathbb{A}$, and is denoted by $\mathsf{var}(\mathbb{A})$.

An operation f on a set D is said to be *idempotent* if the equality $f(x,\dots,x) = x$ holds for all $x \in D$. An algebra all of whose term operations are idempotent is said to be *idempotent*. Recall that δ_d denotes the predicate that is true only on d. It is easily seen that for any constraint language Γ over D all polymorphisms of $\Gamma_{\mathrm{id}} = \Gamma \cup \{\delta_d\colon d \in D\}$ are idempotent. By Theorem 4.4 $\#\mathrm{CSP}(\Gamma)$ and $\#\mathrm{CSP}(\Gamma_{\mathrm{id}})$ are Turing reducible to each other.

An idempotent algebra $\mathbb{A}$ is said to be *congruence singular* if for any its congruences η, θ the matrix $M_{\eta,\theta}$ is block rank-one. If every finite algebra in a variety is congruence singular then the variety is called congruence singular. It can be shown that a constraint language is congruence singular if and only if $\mathsf{Alg}(\Gamma_{\mathrm{id}})$ generates a congruence singular variety. Thus we obtain another equivalent condition for the dichotomy result on unweighted #CSPs.

5. Unweighted #CSP dichotomy

In this section we outline the proof of the following dichotomy theorem for unweighted #CSP.

Theorem 5.1 ([8] and [32, 33]). *Let Γ be a constraint language. Then the following conditions are equivalent:*

(1) $\#\mathrm{CSP}(\Gamma)$ *is solvable in polynomial time;*

(2) Γ *is congruence singular;*

(3) $\mathsf{Alg}(\Gamma_{\mathrm{id}})$ *generates a congruence singular variety;*

(4) Γ *is strongly balanced.*

If none of these conditions hold, $\#\mathrm{CSP}(\Gamma)$ is #P-hard.

The concept central to the proof is that of *compact representation* ([9]) or *frames* ([32, 33]). Let R be an n-ary relation over D with a Mal'tsev polymorphism. For $i \in [n]$ and $a, b \in D$ we write $a \sim_i b$ if there are $\mathbf{a}, \mathbf{b} \in R$ such that $\mathrm{pr}_{[i-1]}\mathbf{a} = \mathrm{pr}_{[i-1]}\mathbf{b}$, and $\mathbf{a}[i] = a$, $\mathbf{b}[i] = b$. Since R is rectangular, the relation $\sim_i$ is an equivalence relation on $\mathrm{pr}_i R$. Let $E_{i1}, \dots, E_{ik_i}$ be the equivalence classes of $\sim_i$. A set $F \subseteq R$ is said to be a compact representation (a frame) of R if

(a) $\mathrm{pr}_i F = \mathrm{pr}_i R$ for each $i \in [n]$; and

(b) for each E_{ik}, $k \in [k_i]$, $i \in [n]$, there is $\mathbf{a}_{ik} \in D^{i-1}$ such that, for each $a \in E_{ik}$, there exists $\mathbf{b}_a \in F$ with $\mathrm{pr}_{[i-1]}\mathbf{b}_a = \mathbf{a}_{ik}$ and $\mathbf{b}_a[i] = a$ (in a compact representation for each pair $(a, b) \in E_{ik}$ there are $\mathbf{b}_a, \mathbf{b}_b$ such that $\mathrm{pr}_{[i-1]}\mathbf{b}_a = \mathrm{pr}_{[i-1]}\mathbf{b}_b$ and $\mathbf{b}_a[i] = a, \mathbf{b}_b[i] = b$).

Lemma 5.2 ([9],[33]).

(1) *If R is an n-ary relation with a Mal'tsev polymorphism φ then (a) there exists its compact representation $F \subseteq R$ such that $|F| \leq n(|D| - 1) + 1$, and (b) R is the* closure *of any its compact representation, that is, every tuple from R can be obtained by (repeatedly) applying φ to tuples from F.*

(2) *If Γ is a constraint language with a Mal'tsev polymorphism, then for any instance $I = R_1(\mathbf{x}_1) \wedge \ldots \wedge R_m(\mathbf{x}_m)$ of $\#\mathrm{CSP}(\Gamma)$ with set of variables V a compact representation F of the set of solutions of I can be found in time polynomial in $|V|$ and m that satisfies the conditions of part (1).*

We now describe the algorithm from [33] to find the number of solutions of an instance $I \in \#\mathrm{CSP}(\Gamma)$ where Γ satisfies the conditions of Theorem 5.1. In particular, by Theorem 4.8 Γ has a Mal'tsev polymorphism φ, hence, by Lemma 5.2 we assume a compact

representation F of the set of solutions of I is known. Thus, the problem is given a compact representation F of an n-ary relation R to find $|R|$.

For $1 \le i < j \le n$ set $N_{i,j}(a) = |\{\mathbf{c}\colon (\mathbf{c}, a) \in \mathrm{pr}_{[i]\cup\{j\}}R\}|$, and let $D_i = \mathrm{pr}_i R$. Since $|R| = \sum_{a\in D_n} N_{n-1,n}(a)$, it suffices to find the numbers $N_{i,j}(a)$. We compute these numbers inductively. First, observe that $N_{1,j}(a)$ is the number of pairs in the binary relation $\mathrm{pr}_{\{1,j\}}R \cap (D \times \{a\})$, and $\mathrm{pr}_{i,j}R$ is the closure of $\mathrm{pr}_{i,j}F$, and therefore can be found by brute force. Thus, it suffices to show how to find $N_{i,j}(a)$ from $N_{i-1,j}$ and $N_{i-1,i}$.

Take particular i and j and suppose that the numbers $N_{i-1,k}$ are found for all $k \ge i$. Let $J = [i] \cup \{j\}$ and let $S = \mathrm{pr}_J R$ viewed as a ternary relation $S = \{(\mathbf{c}, x, y) \in \mathrm{pr}_J R\colon \mathbf{c} \in \mathrm{pr}_{[i-1]}R, x \in D_i, y \in D_j\}$. Since R is strongly balanced, the matrix $M[x, y] = |\{\mathbf{c} \in \mathrm{pr}_{[i-1]}R\colon (\mathbf{c}, x, y) \in S\}|$ is a block rank-one matrix. As $N_{i,j}(y) = \sum_{x\in D_i} M[x, y]$, it suffices to find the entries of M.

Let $S_y(x)$ denote the set $\{\mathbf{c}\colon (\mathbf{c}, x, y) \in S\}$. Since S is rectangular, the relation $\theta_y = \{(x_1, x_2) \in D_i^2\colon S_y(x_1) \cap S_y(x_2) \ne \varnothing\}$ on D_i is an equivalence relation, and $S_y(x_1) = S_y(x_2)$ whenever $(x_1, x_2) \in \theta_y$. Thus if $T(y) \subseteq D_i$ contains one representative of each equivalence class of θ_y, then

$$\sum_{x\in T(y)} M[x, y] = |\{\mathbf{c}\colon \exists x(\mathbf{c}, x, y) \in S\}| = N_{i-1,j}(y). \tag{5.1}$$

We define an equivalence relation θ'_x for $x \in D_i$ in a similar way.

Again using the rectangularity of S, it is not hard to see that if $y, y' \in D_j$ are such that $(x, y), (x, y') \in \mathrm{pr}_{\{i,j\}}S$ for some x, then $\theta_y = \theta_{y'}$. Moreover, in this case $(x, y), (x, y')$ belong to the same block of M. Thus, the equivalence classes of θ_y can be found from the compact representation. The relations $\sim_{i,j}$ and $\sim_{j,i}$, such that $(x_1, x_2) \in \sim_{i,j}$ if and only if $\exists \mathbf{c}, y((\mathbf{c}, x_1, y), (\mathbf{c}, x_2, y) \in S)$, and $(y_1, y_2) \in \sim_{j,i}$ if and only if $\exists \mathbf{c}, x((\mathbf{c}, x, y_1), (\mathbf{c}, x, y_2) \in S)$ are also equivalence relations, and their equivalence classes can also be found from the compact representation F. The matrix M has identical rows corresponding to the equivalence classes of $\sim_{i,j}$, and identical columns corresponding to the equivalence classes of $\sim_{j,i}$.

If $T'(x)$ contains one representative of each of the classes of θ'_x, we have

$$\sum_{y\in T'(x)} M[x, y] = |\{\mathbf{c}\colon \exists y(\mathbf{c}, x, y) \in S\}| = N_{i-1,i}(x). \tag{5.2}$$

The matrix $\widehat{M}$, obtained by choosing one representative from each of the equivalence classes of $\sim_{i,j}$ and $\sim_{j,i}$, is also a block rank-one matrix. Moreover, we know the block structure, row and column sums of $\widehat{M}$ from $\mathrm{pr}_{i,j}R$, $\sim_{i,j}$, $\sim_{j,i}$, (5.1), and (5.2). Hence, we can reconstruct all the entries of $\widehat{M}$. Then, using $\mathrm{pr}_{i,j}R$, $\sim_{i,j}$, and $\sim_{j,i}$, we can reconstruct the matrix M.

Therefore $\#\mathrm{CSP}(\Gamma)$ can be solved in polynomial time.

Somewhat surprisingly the dichotomy result for unweighted #CSPs, Theorem 5.1, can be easily generalized to a dichotomy theorem for weighted #CSPs provided the weights are non-negative rational. Given a collection of functions $\Gamma \subseteq \mathcal{F}(D, \mathbb{Q}^+)$ the idea is to create a new domain for the problem consisting of tuples from D^k, k is the maximum arity of functions in Γ; and then to simulate their weights by introducing several copies of each tuple. Two different ways to implement this approach are suggested in [11].

Theorem 5.3 ([11]). *Let Γ be a finite set of functions from $\mathcal{F}(D, \mathbb{Q}^+)$. Then $\#\mathrm{CSP}(\Gamma)$ is either solvable in polynomial time, or #P-hard.*

Another issue related to dichotomy results is the *metaproblem*: Given a finite set Γ of relations or functions, decide whether or not $\#\mathrm{CSP}(\Gamma)$ can be solved in polynomial time. The metaproblem for unweighted #CSPs was left open even after the dichotomy result [8], because it was not clear if verifying the congruence singularity condition is decidable. The metaproblem for unweighted #CSP has been proved decidable and belonging to NP by Dyer and Richerby [33]. The decidability result can be extended to weighted #CSPs with nonnegative rational weights through the reduction used in Theorem 5.3. However, this reduction does not imply immediately that the metaproblem belongs to NP, because if the weights are large the reduction is inefficient, as it introduces the number of elements into the new domain proportional to the weights of the original CSP.

6. Complex weights and the weighted #CSP dichotomy

In this section we state the dichotomy result by Cai and Chen [16, 17] for arbitrary #CSP with complex weights. Unfortunately, the solution algorithm is too complicated to describe it here. However, we will try to demonstrate how the algorithmic approaches considered before converge in this result.

We start with two results for special cases of #CSP with complex weights that lead up to the general dichotomy. These are generalizations of Theorems 3.9 and 3.10 to complex weights [18, 21].

Theorem 6.1 ([21]). *Let $\Gamma \subseteq \mathcal{F}(\{0,1\},\mathbb{C})$. The problem $\#\mathrm{CSP}(\Gamma)$ is solvable in polynomial time if and only if one of the following two conditions hold:*

(1) *Every function $f \in \Gamma$ has the form $f(\mathbf{x}) = wg(\mathbf{x})i^{L_1(\mathbf{x})+\ldots+L_k(\mathbf{x})}$, where $w \in \mathbb{R}^+$, each L_j is a linear polynomial over $GF(2)$ while addition in the exponent is that of integers, and g is an affine predicate (see Example 2.5);*

(2) *Every function $f \in \Gamma$ can be represented as $f(\mathbf{x}) = h_k(x_1)\ldots h_n(x_n)g(\mathbf{x})$, $\mathbf{x} = (x_1,\ldots,x_n)$, where each h_i is a unary function, and g is a polynomial which is a product of binary polynomials of the form $x_j + x_\ell$ and $x_j + x_\ell + 1$. (Note that in this case f satisfies the block rank-one condition.)*

Otherwise $\#\mathrm{CSP}(\Gamma)$ is #P-hard.

Cai et al. in [18] generalized the results on nonnegative and real binary functions, Theorems 3.4, and Theorem 3.10, respectively, to complex binary functions. More precisely, their result characterizes the complexity of $\mathrm{EVAL}(M)$, where M is either symmetric or is of the form $\mathsf{bip}'(A) = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}$ for some matrix A (we will call such matrices *bipartite*). This result, first, requires a generalization of Hadamard matrices, called discrete unitary matrices; and, second, a more general solution algorithm for discrete unitary matrices.

The first step is to represent an arbitrary complex $n \times n$-matrix M in a special form. The problem $\mathrm{EVAL}(M)$ for an arbitrary symmetric or bipartite matrix M is either #P-hard, or is Turing interreducible with $\mathrm{EVAL}(M')$ for a *purified* matrix $M' = \mathsf{bip}'(A)$, for A of the

following form

$$A = \begin{pmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_k \end{pmatrix} \begin{pmatrix} \xi_{11} & \xi_{12} & \dots & \xi_{1n-k} \\ \xi_{21} & \xi_{22} & \dots & \xi_{2n-k} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{k1} & \xi_{k2} & \dots & \xi_{kn-k} \end{pmatrix} \begin{pmatrix} \mu_{k+1} & & & \\ & \mu_{k+2} & & \\ & & \ddots & \\ & & & \mu_n \end{pmatrix},$$

for some $1 \le k \le n$, where $\mu_j > 0$ and every ξ_{ij} is a root of unity.

The next step proves that $\mathrm{EVAL}(M)$ is either #P-hard, or is interreducible with $\mathrm{EVAL}(M')$ with vertex weights, where $M' = \mathsf{bip}'(A)$ for a *discrete unitary matrix* A, that is, an $m \times m$-matrix A satisfying the following conditions: (a) Every entry of A is a root of unity, (b) $A(1,i) = A(j,1) = 1$ for $i, j \in [m]$, and (c) AA^* and A^*A are diagonal matrices (A^* denotes the conjugate transpose). Finally, let ζ_q denote a qth primitive root of unity and let F_q be a discrete unitary matrix given by $F_q[x,y] = \zeta_q^{(x-1)(y-1)}$.

Theorem 6.2 ([18]). *Let $M = \mathsf{bip}'(A)$ for a discrete unitary matrix A. Then* $\mathrm{EVAL}(M)$ *is solvable in polynomial time if and only if A is a* Kronecker product $F_{q_1} \otimes \ldots \otimes F_{q_\ell}$ *for some $q_1, \ldots, q_\ell$. Otherwise* $\mathrm{EVAL}(M)$ *is #P-hard.*

The solution algorithm for the weighted #CSP over F_q for $q = p^k$ boils down to computing sums of the form

$$\sum_{x_1,\ldots,x_n \in \mathbb{Z}_q} \zeta_q^{s(x_1,\ldots,x_n)},$$

where s is a quadratic polynomial over $GF(q)$.

We now turn to a description of the dichotomy result for complex weighted #CSPs obtained in [16, 17]. Let $\Gamma \subseteq \mathcal{F}(D, \mathbb{C})$. The result we are going to state requires 3 properties of Γ that generalize the properties of functions used so far.

Block orthogonality. A function $f \in \mathcal{F}(D, \mathbb{C})$, $D = \{d_1, \ldots, d_\ell\}$, is said to satisfy the absolute block rank-one condition if $|f|$, that is, the function that takes the absolute value of f, satisfies the block rank-one condition. Let $\mathbf{x}, \mathbf{y} \in \mathbb{C}^\ell$ be vectors that are linearly dependent after taking absolute values. In particular, they have the same nonzero entries; denote the set of such entries T. Let $T_1, \ldots, T_k$ be the partition of T defined by $|\mathbf{x}[i]| = |\mathbf{x}[j]|$ if and only if $i, j \in T_s$ for some s. Clearly, if we start with $\mathbf{y}$ the partition will be the same. Vectors $\mathbf{x}, \mathbf{y}$ are called *block-orthogonal* if for every $s \in [k]$ they satisfy $\sum_{i \in T_s} \mathbf{x}[i] \cdot \overline{\mathbf{y}[i]} = 0$. Let $f(\mathbf{x}, *)$ denote the ℓ-ary vector $(f(\mathbf{x}, d_1), \ldots, f(\mathbf{x}, d_\ell))$. An absolute block rank-one function $f : D^n \to \mathbb{C}$ is said to be *block-orthogonal* if for any $\mathbf{x}, \mathbf{y} \in D^{n-1}$ such that $f(\mathbf{x}, *), f(\mathbf{y}, *) \neq 0$ and vectors $|f(\mathbf{x}, *)|, |f(\mathbf{y}, *)|$ are linearly dependent, the vectors $f(\mathbf{x}, *), f(\mathbf{y}, *)$ are either linearly dependent, or block-orthogonal.

Type partition. Let $\Phi \subseteq D^n$ be a nonempty set and $\Psi_1, \ldots, \Psi_k$ be a partition of Φ, $k \ge 1$. These sets define a mapping $\mathsf{type}(\cdot)$ from $D^1 \cup \ldots \cup D^n$ to $[k]$ as follows: for $\ell \in [n]$ and $\mathbf{x} \in D^\ell$, we set $\mathsf{type}(\mathbf{x}) = \{j \in [k] \colon \exists \mathbf{y} \in \Psi_j$ such that $\mathbf{x} = \mathrm{pr}_{[\ell]}\mathbf{y}\} \subseteq [k]$. The mapping $\mathsf{type}(\cdot)$ is said to be a *type-partition* map if for any $\ell \in [n]$ and $\mathbf{x}, \mathbf{y} \in D^\ell$, the sets $\mathsf{type}(\mathbf{x})$ and $\mathsf{type}(\mathbf{y})$ are either equal or disjoint.

Let $f : D^n \to \mathbb{C}$. Also, let $\Phi = D^{n-1}$ and sets $\Psi_i \subseteq \Phi$ be defined by the rule: $\mathbf{x}, \mathbf{y} \in \Psi_i$

for some i if and only if $f(\mathbf{x}, *), f(\mathbf{y}, *)$ are linearly dependent. The corresponding mapping is denoted by type_f, and the corresponding equivalence relation on D^{n-1} by θ_f.

Mal'tsev condition. For a function $f : D^n \to \mathbb{C}$ let $\mathsf{supp}(f) \subseteq D^n$ denote the relation that contains all $\mathbf{x} \in D^n$ such that $f(\mathbf{x}) \neq 0$. Let $\Gamma \subseteq \mathcal{F}(D, \mathbb{C})$. It is said to satisfy the Mal'tsev condition if there is a Mal'tsev operation φ such that for any (n-ary) function f pp-definable in Γ the operation φ is a polymorphism of $\mathsf{supp}(f)$ and θ_f, where the latter is viewed as the set of all $2n-2$-tuples $(\mathbf{x}, \mathbf{y})$ whenever $\mathbf{x}, \mathbf{y}$ are θ_f-related.

Theorem 6.3 ([17]). *Let $\Gamma \subseteq \mathcal{F}(D, \mathbb{C})$. Then $\#\mathrm{CSP}(\Gamma)$ is solvable in polynomial time if and only if Γ satisfies the Mal'tsev condition, every function f pp-definable in Γ is absolute block rank-one, block-orthogonal, and type_f is a type-partition map. Otherwise $\#\mathrm{CSP}(\Gamma)$ is #P-hard.*

7. Conclusion

This short survey is inevitably biased and incomplete. We conclude with a short overview of what is left out, some open problems, and potential future research directions.

The study of general counting and weighted #CSPs may be considered as mostly completed with only some side issues remaining. The most important of them is the metaproblem: Given a constraint language or a set of functions, decide whether the corresponding #CSP is solvable in polynomial time, or, more generally, determine the complexity of the corresponding #CSP. While the complexity of the metaproblem for unweighted #CSPs is to some extent understood, that for weighted #CSPs remains open.

A different direction in the study of the CSP is to allow restricted classes of inputs, such as CSPs on graphs or relational structures of bounded degree, or planar, or bipartite, etc. Although a number of results have been obtained for such problems, see, e.g. [25–28, 39, 42, 60, 64], the complexity of these problems remains unknown for many important classes of structures. For example, any problem $\#\mathrm{CSP}(\Gamma)$ is solvable in polynomial time on structures of bounded tree-width [24]. A compelling question is then what is the trade-off between the constraint language Γ and the class of allowed inputs so that the corresponding #CSP remains in FP.

We also did not touch another research direction related to counting CSPs, the complexity of computing the Holant problem. This problem originated from the work of Valiant on holographic algorithms [63, 64]. The Holant problem is somewhat more expressive than the CSP, however, its complexity is not completely known even in relatively small cases; for some of the existing results on the subject see [19, 20]. Also, both the #CSP and Holant can be considered to count solutions modulo a certain integer, for instance, counting the parity of the number of solutions. Although some results in this direction exists, the problem mostly remains open.

Finally, the complexity of approximation of all counting problems has been studied. The complexity landscape in this case is much more complicated, and many of the natural problems remain open.

References

[1] Allender, E. and Ogihara, M, *Relationships among PL, #L, and the determinant*, In Structure in Complexity Theory Conference, pp. 267–278, 1994.

[2] Barto, L. and Kozik, M, *Constraint satisfaction problems of bounded width*, In FOCS, 2009.

[3] Beigel, R, *Relativized counting classes: Relations among thresholds, parity, and mods*, J. Comput. Syst. Sci., **42**(1) (1991), 76–96.

[4] _____, *Perceptrons, PP, and the polynomial hierarchy*, Computational Complexity, **4** (1994), 339–349.

[5] Bodnarchuk, V.G., Kaluzhnin, L.A., Kotov, V.N., and Romov, B.A, *Galois theory for Post algebras. I*, Kibernetika, **3** (1969), 1–10.

[6] Bulatov, A.A, *Complexity of conservative constraint satisfaction problems*, ACM Trans. Comput. Log., **12**(4) (2011), 24.

[7] Bulatov, A.A, *A dichotomy theorem for constraint satisfaction problems on a 3-element set*, J. ACM, **53**(1) (2006), 66–120.

[8] _____, *The complexity of the counting constraint satisfaction problem*, J. ACM, **60**(5) (2013), 34.

[9] Bulatov, A.A. and Dalmau, V, *A simple algorithm for Mal'tsev constraints*, SIAM J. Comput., **36**(1) (2006), 16–27.

[10] Bulatov, A.A. and Dalmau, V, *Towards a dichotomy theorem for the counting constraint satisfaction problem*, Inf. and Comp., **205**(5) (2007), 651–678.

[11] Bulatov, A.A., Dyer, M.E., Goldberg, L.A., Jalsenius, M., Jerrum, M., and Richerby, D, *The complexity of weighted and unweighted #CSP*, J. Comput. Syst. Sci., **78**(2) (2012), 681–688.

[12] Bulatov, A.A., Dyer, M.E., Goldberg, L.A., Jalsenius, M., and Richerby, D, *The complexity of weighted Boolean #CSP with mixed signs*, Theor. Comput. Sci., **410**(38-40) (2009), 3949–3961.

[13] Bulatov, A.A., Dyer, M.E., Goldberg, L.A., Jerrum, M., and McQuillan, C, *The expressibility of functions on the Boolean domain, with applications to counting CSPs*, J. ACM, **60**(5) (2013), 32.

[14] Bulatov, A.A. and Grohe, M, *The complexity of partition functions*, Theor. Comput. Sci., **348**(2-3) (2005), 148–186.

[15] Bulatov, A.A., Jeavons, P., and Krokhin, A.A, *Classifying the complexity of constraints using finite algebras*, SIAM J. Comput., **34**(3) (2005), 720–742.

[16] Cai, J.-Y., and Chen, X, *Complexity of counting CSP with complex weights*, CoRR, arXiv:1111.2384, 2011.

[17] Cai, J.-Y., and Chen, X, *Complexity of counting CSP with complex weights*, In STOC, pp. 909–920, 2012.

[18] Cai, J.-Y., Chen. X., and Lu, P, *Graph homomorphisms with complex values: A dichotomy theorem*, SIAM J. Comput., **42**(3) (2013), 924–1029.

[19] Cai, J.-Y., Huang, S., and Lu, P, *From Holant to #CSP and back: Dichotomy for Holant-c problems*, Algorithmica, **64**(3) (2012), 511–533.

[20] Cai, J.-Y., Lu, P, and Xia, M., *Dichotomy for Holant* problems with domain size 3*, In SODA, pp. 1278–1295, 2013.

[21] _____, *The complexity of complex weighted Boolean #CSP*, J. Comput. Syst. Sci., **80**(1) (2014), 217–236.

[22] Creignou, N., and Hermann, M, *Complexity of generalized satisfiability counting problems*, Inf. and Comp., **125**(1) (1996), 1–12.

[23] Creignou, N., Khanna, S., and Sudan, M, *Complexity Classifications of Boolean Constraint Satisfaction Problems*, volume 7 of *SIAM Monographs on Discrete Mathematics and Applications*, SIAM, 2001.

[24] Dalmau, V., and Jonsson, P. *The complexity of counting homomorphisms seen from the other side*, Theor. Comput. Sci., **329**(1-3) (2004), 315–323.

[25] Díiaz, J., Serna, M.J., and Spirakis, P. G., *On the random generation and counting of matchings in dense graphs*, Theor. Comput. Sci., **201**(1-2) (1998), 281–290.

[26] Díiaz, J., Serna, M.J., and Thilikos, D.M, *Counting H-colorings of partial k-trees*, Theor. Comput. Sci., **281** (2002), 291–309.

[27] _____, *Fixed parameter algorithms for counting and deciding bounded restrictive list H-colorings*, In ESA, pp. 275–286, 2004.

[28] Dyer, M.E., Frieze, A., and Jerrum, M., *On counting independent sets in sparse graphs*, SIAM J. on Comput., **31** (2002), 1527–1541.

[29] Dyer, M.E., Goldberg, L.A., and Paterson, M, *On counting homomorphisms to directed acyclic graphs*, In ICALP, pp. 38–49, 2006.

[30] Dyer, M.E. and Greenhill, C. *The complexity of counting graph homomorphisms*, Random Structures and Algorithms, **17** (2000), 260–289.

[31] Dyer, M.E., Goldberg, L.A., and Jerrum, M., *A complexity dichotomy for hypergraph partition functions*, Comput. Compl., **19**(4) (2010), 605–633.

[32] Dyer, M.E. and Richerby, D, *On the complexity of #CSP*, In STOC, pp. 725–734, 2010.

[33] _____, *An effective dichotomy for the counting constraint satisfaction problem*, SIAM J. Comput., **42**(3) (2013), 1245–1274.

[34] Feder, T. and Vardi, M.Y, *The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory*, SIAM J. Comput., **28** (1998), 57–104.

[35] Geiger, D, *Closed systems of function and predicates*, Pacific J. Math., pp. 95–100, 1968.

[36] Goldberg, L.A., Grohe, M., Jerrum, M., and Thurley, M., *A complexity dichotomy for partition functions with mixed signs*, SIAM J. Comput., **39**(7) (2010), 3336–3402.

[37] Grätzer, G, *Universal algebra*, Springer, 2nd edition, 2008.

[38] Green, F., Köbler, J., Regan, K.W., Schwentick, T., and Torán, J, *The power of the middle bit of a #P function*, J. Comput. Syst. Sci., **50**(3) (1995), 456–467.

[39] Greenhill, C, *The complexity of counting colourings and independent sets in sparse graphs and hypergraphs*, Comput. Compl., **9** (2000), 52–73.

[40] Hell, P., and Nešetřil, J, *On the complexity of H-coloring*, J. Comb. Theor., Ser. B, **48** (1990), 92–110.

[41] Hemaspaandra, L.A. and Vollmer, H, *The satanic notations: counting classes beyond #P and other definitional adventures*, SIGACT News, **26**(1) (1995), 2–13.

[42] Hunt III, H.B., Marathe, M.V., Radhakrishnan, V., and Stearns, R.E, *The complexity of planar counting problems*, SIAM J. Comput., **27** (1998), 1142–1167.

[43] Idziak, P., Markovic, P., McKenzie, R., Valeriote, M., and Willard, R, *Tractability and learnability arising from algebras with few subpowers*, In LICS, pp. 213–224, 2007.

[44] Ising, E, *Beitrag zur theorie des ferromagnetismus*, Zeitschrift fur Physik, **31** (1925), 253–258.

[45] Jeavons, P.G., Cohen, D.A., and Gyssens, M, *Closure properties of constraints* J. ACM, **44** (1997), 527–548.

[46] Jerrum, M. and Sinclair, A, *Polynomial-time approximation algorithms for the Ising model*, SIAM J. Comput., **22**(5) (1993), 1087–1116.

[47] Ko, K.-I, *Complexity theory of real functions*, Birkhäuser, 1990.

[48] Ladner, R, *On the structure of polynomial time reducibility*, J. ACM, **22** (1975), 155–171.

[49] Levin, L.A, *Universal enumeration problems*, Probl. Inf. Transm., **9** (1973), 265–266.

[50] Lidl, R. and Niederreiter, H, *Finite fields*, volume 20 of Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2 edition, 1997.

[51] Lovász, L, *Operations with structures*, Acta. Math. Acad. Sci. Hung., **18** (1967), 321–328.

[52] Nordh, G. and Jonsson, P., *The complexity of counting solutions to systems of equations over finite semigroups*, In COCOON, pp. 370–379, 2004.

[53] Potts, R. *Some generalized order-disorder transformations*, Proc. Cambridge Philos. Soc., **48** (1952), 106–109.

[54] Provan, J.S. and Ball, M.O, *The complexity of counting cuts and of computing the probability that a graph is connected*, SIAM J. Comput., **12**(4) (1983), 777–788.

[55] Saluja, S., Subrahmanyam, K.V., and Thakur, M.N, *Descriptive complexity of #P functions*, J. Comput. Syst. Sci., **50**(3) (1995), 493–505.

[56] Schaefer, T.J, *The complexity of satisfiability problems*, In STOC, pp. 216–226, 1978.

[57] Schöning, U, *The power of counting*, In Complexity Theory Retrospective, Springer, 1990, pp. 204–223.

[58] Thurley, M, *The complexity of partition functions on Hermitian matrices*, CoRR, arXiv: 1004.0992, 2010.

[59] Toda, S. and Ogiwara, M, *Counting classes are at least as hard as the polynomial-time hierarchy*, SIAM J. Comput., **21**(2) (1992), 316–328.

[60] Vadhan, S.P, *The complexity of counting in sparse, regular and planar graphs*, SIAM J. Comput., **31**(2) (2001), 398–427.

[61] Valiant, L, *The complexity of computing the permanent*, Th. Comput. Sci., **8** (1979), 189–201.

[62] _____, *The complexity of enumeration and reliability problems*, SIAM J. Comput., **8**(3) (1979), 410–421.

[63] _____, *Expressiveness of matchgates*, Th. Comput. Sci., **289**(1) (2002), 457–471.

[64] _____, *Holographic algorithms (extended abstract)*, In FOCS, pp. 306–315, 2004.

School of Computing Science, Simon Fraser University, Burnaby, Canada
E-mail: abulatov@sfu.ca

Flows, cuts and integral routing in graphs - an approximation algorithmist's perspective

Julia Chuzhoy

Abstract. Flow, cut and integral graph routing problems are among the most extensively studied in Operations Research, Optimization, Graph Theory and Computer Science. We survey known algorithmic results for these problems, including classical results and more recent developments, and discuss the major remaining open problems, with an emphasis on approximation algorithms.

Mathematics Subject Classification (2010). Primary 68Q25; Secondary 68Q17, 68R05, 68R10.

Keywords. Maximum flow, minimum cut, network routing, approximation algorithms, hardness of approximation, graph theory.

1. Introduction

In this survey we consider flow, cut, and integral routing problems in graphs. These three types of problems are among the most extensively studied in Operations Research, Optimization, Graph Theory, and Computer Science. Problems of these types naturally arise in many applications, and algorithms for solving them are among the most valuable and powerful tools in algorithm design and analysis.

In the classical maximum s–t flow problem, we are given an n-vertex graph $G = (V, E)$, that can be either directed or undirected, with non-negative capacities $c(e)$ on edges $e \in E$, and two special vertices: s, called the source, and t, called the destination. Let $\mathcal{P}$ be the set of all paths connecting s to t in G. An s–t flow f is an assignment of non-negative values $f(P)$ to all paths $P \in \mathcal{P}$, such that for each edge $e \in E$, the flow through e does not exceed its capacity $c(e)$, that is, $\sum_{P:e\in P} f(P) \leq c(e)$. The value of the flow f is $\sum_{P\in\mathcal{P}} f(P)$, and the goal is to find a flow of maximum value. The maximum flow problem was introduced in the 50's in order to model the capacity of the Soviet and East European railway systems. Ford and Fulkerson [41] were the first to provide an efficient algorithm for solving the problem. The problem can be expressed as a linear program (LP):

$$\begin{array}{lll} \text{(LP-flow)} & \max & \sum_{P\in\mathcal{P}} f(P) \\ & \text{s.t.} & \\ & & \sum_{P:e\in P} f(P) \leq c(e) \quad \forall e \in E \\ & & f(P) \geq 0 \qquad\quad \forall P \in \mathcal{P} \end{array}$$

So far, in our definition of the maximum s–t flow problem, the number of paths P with non-zero flow value $f(P)$ may be exponentially large in the graph size, and so can the num-

Proceedings of the International Congress of Mathematicians, Seoul, 2014

ber of variables of (LP-flow). Fortunately, there is an equivalent "compact" LP-formulation of the problem, whose solution can be efficiently converted into a solution to (LP-flow), where the number of paths P with $f(P) > 0$ is bounded by $|E|$. This provides an efficient algorithm to solve (LP-flow), as long as we are only required to list the non-zero values $f(P)$ in the solution.

A very useful feature of the maximum s-t flow problem is that, if all edge capacities $c(e)$ are integral, then there is a maximum flow where for each $P \in \mathcal{P}$, $f(P)$ is integral, and such a flow can be found efficiently. This property is often referred to as the *integrality of flow*. In particular, if all edge capacities are unit, then we can efficiently find a maximum-cardinality collection $\mathcal{P}'$ of paths connecting s to t, such that the paths in $\mathcal{P}'$ are *edge-disjoint*: that is, every edge of G belongs to at most one path of $\mathcal{P}'$.

A problem closely related to maximum s–t flow is minimum s–t cut. The input to this problem is the same as the input to the maximum s–t flow problem, only now we will think of the values $c(e)$ as edge costs and not capacities. The goal is to select a minimum-cost subset $E' \subseteq E(G)$ of edges, such that $G \setminus E'$ contains no path connecting s to t, where the cost of E' is $\sum_{e \in E'} c(e)$. It is easy to see that the value of the maximum s–t flow cannot exceed the value of the minimum s–t cut in any graph: every path $P \in \mathcal{P}$ must contain at least one edge of E', and so the total flow carried by the paths in $\mathcal{P}$ cannot exceed the total capacity of the edges in E'. The classical result of Ford and Fulkerson [41], often referred to as the Max-Flow Min-Cut Theorem, shows that the opposite is also true, that is, in any graph, the value of the minimum s–t cut is equal to the value of the maximum s–t-flow! In fact, their algorithm for computing maximum flow can also be used to compute a minimum cut. Therefore, we can see minimum cut as revealing the bottleneck in the routing capacity of a graph: if the maximum amount of flow that can be sent from s to t is x, then we can produce a certificate for this fact in the form of a valid flow of value x, and an s–t cut of cost x. A convenient way of seeing the connection between flows and cuts is by computing the dual linear program of (LP-flow), that we will call (LP-cut) for reasons that will become apparent below.

$$\begin{aligned}
\text{(LP-cut)} \qquad & \min \quad \textstyle\sum_{e \in E} c_e x_e & & \\
& \text{s.t.} & & \\
& \textstyle\sum_{e \in P} x_e \geq 1 & \forall P \in \mathcal{P} & (1.1)\\
& x_e \geq 0 & \forall e \in E & (1.2)
\end{aligned}$$

Even though the number of constraints in (LP-cut) may be exponential in the graph size, it can still be solved efficiently by standard methods, such as the Ellipsoid algorithm with a separation oracle.

Let us start by adding the following *integrality* constraints to (LP-cut):

$$x_e \in \{0, 1\} \qquad \forall e \in E \tag{1.3}$$

This combination of a linear program with integrality constraints is called *integral linear program*, and we denote it by (ILP-cut). It is immediate to see that (ILP-cut) is equivalent to the minimum s–t cut problem: we set $x_e = 1$ if e belongs to the solution E', and $x_e = 0$ otherwise. Constraint (1.1) ensures that every path from s to t contains at least one edge from E' - that is, $G \setminus E'$ contains no s–t path. Of course, any feasible solution of (ILP-cut) is also a feasible solution to (LP-cut). However, (LP-cut) allows more solutions: for example, solutions where the variables x_e take fractional values. We say that (LP-cut) is a *relaxation*

of the minimum s–t cut problem. The optimal solution to (LP-cut) is called the optimal *fractional* solution to minimum s–t cut, and its value is denoted by $\mathsf{OPT_{LP}}$. The optimal solution to (ILP-cut) is called the optimal *integral* solution; its value is denoted by OPT, and it is also the value of the minimum s–t cut in G. Since the optimal integral solution is a valid solution to (LP-cut), $\mathsf{OPT_{LP}} \leq \mathsf{OPT}$ must hold. Interestingly, in this particular linear program, the two values are equal. Moreover, there is an efficient algorithm, that, given a fractional solution to (LP-cut), computes a valid integral solution of the same value.

We describe the algorithm for directed graphs; the algorithm for undirected graphs is similar, with minor adjustments. The idea is to view the values x_e in the optimal solution of (LP-cut) as edge lengths. We can then define, for every pair (u, v) of vertices, the distance $d(u, v)$ from u to v, to be the length of the shortest path connecting u to v, under the edge lengths x_e. Constraint (1.1) ensures that $d(s, t) \geq 1$, and it is easy to see that for every edge $e = (u, v)$:

$$d(s, v) \leq d(s, u) + x_e \tag{1.4}$$

Let us choose a value $\rho \in (0, 1)$ uniformly at random, and let $B(s, \rho) = \{v \mid d(s, v) \leq \rho\}$ be the ball of radius ρ around s. This ball defines an s–t cut $E'_\rho \subseteq E$, where $e = (u, v) \in E'_\rho$ if $u \in B(s, \rho)$ and $v \notin B(s, \rho)$. The probability that $e = (u, v)$ belongs to E'_ρ is the probability that ρ lies between $d(s, u)$ and $d(s, v)$, which, from (1.4), is bounded by x_e. The expected cost of the cut E'_ρ is then:

$$\sum_{e \in E} c(e) \cdot \mathbf{Pr}_\rho \left[e \in E'_\rho\right] \leq \sum_{e \in E} c(e) x_e = \mathsf{OPT_{LP}}.$$

At least one value $\rho : 0 < \rho < 1$ must satisfy $\sum_{e \in E'_\rho} c_e \leq \mathsf{OPT_{LP}}$. We can find this value by going over all possible values of ρ and computing E'_ρ for each of them. Fortunately, the number of different values of ρ that we need to check is not very large - it is enough to consider all values in set $\{d(s, v) \mid v \in V\}$.

An algorithm that, given a fractional solution to a linear program, computes an integral solution is called an *LP-rounding algorithm*. If the value of the solution produced by the algorithm equals to the value of the fractional solution, then this algorithm can be used to solve the problem exactly. If additionally the LP-rounding algorithm is efficient, then we obtain an efficient algorithm for the problem, thus proving that it is in **P**. This is exactly what we have just shown for minimum s–t cut. However, many optimization problems that we consider in this survey are **NP**-hard, and therefore we do not expect them to have efficient algorithms. Instead, we will often look for *approximation algorithms* - efficient algorithms that solve the problem approximately. Given a minimization problem Π, we say that an efficient algorithm $\mathcal{A}$ is an α-approximation algorithm for Π, if for any instance I of Π, algorithm $\mathcal{A}$ produces a solution of value at most $\alpha \cdot \mathsf{OPT}(I)$, where $\mathsf{OPT}(I)$ is the value of the optimal solution for I. For a maximization problem, an α-approximation algorithm needs to produce a solution of value at least $\mathsf{OPT}(I)/\alpha$. Different optimization problems often have different approximation factors achievable by efficient algorithms. The approximation factor α may be a constant, or some function of the input size n (like $O(\log n)$, $O(\sqrt{n})$, and so on). For many optimization problems, we still do not know what is the best approximation factor α^* achievable for them. In order to determine this factor, in addition to designing an approximation algorithm, that establishes an upper bound on α^*, we need to provide a lower bound on α^*. This is usually done by proving *hardness of approximation*, or *inapproximability* results: namely, that achieving a better than α^*-approximation for a given problem Π is an **NP**-hard problem.

As we have shown, there is an efficient LP-rounding algorithm for minimum s–t cut that can be used, together with (LP-cut), to solve the problem exactly. For many other minimization problems, the value of the integral solution produced by an LP-rounding algorithm for a minimization problem is greater than $\mathsf{OPT}_{\mathsf{LP}}$. However, if the value of the solution is at most $\alpha \cdot \mathsf{OPT}_{\mathsf{LP}}$ for any input instance I, then, since $\mathsf{OPT}_{\mathsf{LP}} \leq \mathsf{OPT}$ we obtain an α-approximate LP-rounding algorithm. The technique of rounding linear programming relaxations is one of the most powerful and widely used tools in the design of approximation algorithms.

From the strong duality theorem, the optimal value of (LP-flow) equals to the optimum value of (LP-cut), that is, maximum flow equals to the value of the minimum fractional cut. But since the values of the optimal fractional and the optimal integral solutions to the s–t cut problem are the same, we get that the maximum flow value equals to the value of the minimum cut in any graph G.

Linear program (LP-cut) is one of the rare cases where the optimal fractional and the optimal integral solutions have the same value. For many other minimization problems and their linear programming relaxations, $\mathsf{OPT}_{\mathsf{LP}} < \mathsf{OPT}$ holds. Given a minimization problem Π, and a linear programming relaxation (LP-rel) for Π, the *integrality gap* of (LP-rel) is the largest possible ratio between the value OPT of the optimal integral solution and the value $\mathsf{OPT}_{\mathsf{LP}}$ of the optimal fractional solution, achieved by any instance I of Π. (For maximization problems we reverse the ratio, and the integrality gap is the maximum of $\mathsf{OPT}_{\mathsf{LP}}/\mathsf{OPT}$ over all instances; so the integrality gap is always at least 1). If the integrality gap of a linear programming relaxation (LP-rel) of problem Π is α, then no LP-rounding algorithm can achieve a better than α-approximation for the problem. This statement however is only true for the specific linear programming relaxation (LP-rel) of Π. Often one can come up with different linear programming relaxations of the same problem, that have different integrality gaps, and LP-rounding algorithms achieving different approximation factors. Studying integrality gaps of linear programs is therefore crucial in understanding the power and the limitations of the LP-rounding approach for specific optimization problems. Often, instances exhibiting large LP-integrality gaps can give us insight into the structure of hard instances of the problem, and this can help us prove inapproximability results. Alternatively, they can help us strengthen the LP relaxation and obtain better LP-rounding algorithms.

As we have already seen, the integrality ratio of (LP-cut) is 1. We have also already mentioned that, if all edge capacities are integral, then there is an optimal solution to the maximum s–t flow problem where all values $f(P)$ are also integral. Therefore, if all edge capacities are integral, then the integrality gap of (LP-flow) is also 1. In the following sections we consider generalizations of the maximum s–t flow problem, where instead of one source-destination pair, there are several such pairs. There are two natural ways to define the objective function in this setting: we can try to maximize the total amount of flow sent between all source-destination pairs - a problem known as the maximum multicommodity flow; or we can try to maximize a value λ such that all demand pairs can simultaneously send λ flow units between them - this is known as maximum concurrent flow. We define the two corresponding graph cut problems, minimum multicut and sparsest cut, and study their LP-relaxations, as well as known approximation algorithms and hardness results in Sections 2 and 3. Unfortunately, the integrality of flow does not hold anymore in the multiple source-destination pairs setting, and the problem of computing maximum integral flow becomes **NP**-hard. We discuss approximation algorithms and hardness results for integral routing problems in Sections 4 and 5.

Before we proceed, let us mention another common and useful version of the maximum

s–t flow problem, where the capacities are on the graph vertices and not on edges. In this problem, we are given a graph $G = (V, E)$, a source vertex $s \in V$ and a destination vertex $t \in V$, and capacity values $c(v)$ for all vertices $v \in V \setminus \{s, t\}$. As before, let $\mathcal{P}$ denote the set of all paths connecting s to t in G. A valid flow assigns values $f(P) \in \mathbb{R}^+$ to each path $P \in \mathcal{P}$, so that for every vertex $v \in V \setminus \{s, t\}$, $\sum_{P:v\in P} f(P) \leq c(v)$. In the corresponding vertex cut problem, we are given costs $c(v)$ on vertices $v \in V$, and the goal is to select a minimum-cost subset $S \subseteq V \setminus \{s, t\}$ of vertices, so that $G \setminus S$ contains no path connecting s to t. The node-capacitated version of the maximum flow problem behaves very similarly to the edge-capacitated one. We can write a linear programming relaxation, similar to (LP-flow), which can be solved efficiently using similar methods. The dual of this linear program is a relaxation of the minimum vertex cut problem. As in the edge-capacitated version of the problem, the integrality gap of the LP-relaxation for minimum vertex cut is 1, and, when all vertex capacities are integral, the integrality gap of the LP-relaxation for node-capacitated maximum flow is also 1. The maximum flow value and the minimum cut value are therefore equal for any graph G even in this model. If all vertex capacities are unit, then we obtain an efficient algorithm for computing a maximum-cardinality set $\mathcal{P}'$ of internally node-disjoint paths connecting s to t (so every vertex $v \in V \setminus \{s, t\}$ may belong to at most one path of $\mathcal{P}'$). Therefore, the cardinality of $\mathcal{P}'$ is equal to the value of the minimum vertex s–t cut in any graph G - this is known as Menger's theorem [69].

2. Maximum multicommodity flow and minimum multicut

A natural generalization of the maximum s–t flow problem is maximum multicommodity flow. In this problem, instead of a single source-destination pair (s, t), we are given a collection of k such pairs $\{(s_1, t_1), \ldots, (s_k, t_k)\}$, that we call *demand pairs*. The goal is to send maximum amount of flow between the demand pairs, without violating the edge capacities: that is, for each $1 \leq i \leq k$, the flow leaving s_i must arrive at t_i, and the total amount of flow traversing any edge e is at most $c(e)$. It is sometimes convenient to think of having k different flow types, or *commodities*, where the ith commodity needs to be sent from s_i to t_i. For each $1 \leq i \leq k$, let $\mathcal{P}_i$ denote the set of all paths connecting s_i to t_i in G. The following linear program is a generalization of (LP-flow) to the multi-commodity setting:

$$\text{(LP-multi-flow)} \qquad \begin{aligned} \max \quad & \textstyle\sum_{i=1}^{k} \sum_{P\in\mathcal{P}_i} f(P) \\ \text{s.t.} \quad & \\ & \textstyle\sum_{P:e\in P} f(P) \leq c(e) \quad \forall e \in E \\ & f(P) \geq 0 \qquad\quad \forall 1 \leq i \leq k \quad \forall P \in \mathcal{P}_i \end{aligned}$$

Like (LP-flow), this linear program can be solved efficiently using similar methods. The cut counterpart of maximum multicommodity flow is minimum multicut. In this problem, the input is the same as in the maximum multicommodity flow problem, but we view the values $c(e)$ as edge costs, rather than capacities. The goal is to select a minimum-cost subset $E' \subseteq E$ of edges, such that in graph $G \setminus E'$, there is no path connecting any source s_i to its destination t_i. As in the single-commodity scenario, it is easy to see that the value of the maximum multicommodity flow cannot exceed the value of the minimum multicut in any graph G, since for each $1 \leq i \leq k$, every path $P \in \mathcal{P}_i$ must contain at least one edge of E'. Therefore, the total amount of flow carried by the paths in $\bigcup_{i=1}^{k} \mathcal{P}_i$ is bounded by the total

capacity of the edges in E'. The dual linear program of (LP-multi-flow) also happens to be a relaxation of minimum multicut:

$$\begin{aligned}
\text{(LP-multicut)} \qquad \min \quad & \textstyle\sum_{e\in E} c(e)x_e & & & \\
\text{s.t.} \quad & & & & \\
& \textstyle\sum_{e\in P} x_e \geq 1 & \forall 1 \leq i \leq k & \forall P \in \mathcal{P}_i & (2.1)\\
& x_e \geq 0 & \forall e \in E & & (2.2)
\end{aligned}$$

Indeed, if we restrict the values x_e to be in $\{0,1\}$, and let E' be the set of all edges e with $x_e = 1$, then Constraint (2.1) ensures that every path connecting any source s_i to its destination t_i contains at least one edge of E', or, equivalently, $G \setminus E'$ contains no path connecting s_i to t_i, for any $1 \leq i \leq k$. Let $\mathsf{OPT}_{\mathsf{LP}}$ be the value of the optimal solution to (LP-multicut), that we also call the *minimum fractional multicut value*. The optimal solution to the minimum multicut problem is denoted by OPT, and is called the *minimum integral multicut value*. From the LP-duality theorem, the value of the maximum multicommodity flow equals to the value of the minimum fractional multicut. However, the integrality gap of (LP-multicut) is no longer 1, and so the maximum multicommodity flow value may be smaller than the value of minimum multicut. The equality between maximum flow and minimum cut therefore breaks down in the multicommodity setting. However, we can still hope to obtain an approximate version of the Max-Flow Min-Cut Theorem, by bounding what is called the *flow-cut gap* - the largest possible ratio between maximum multicommodity flow and minimum multicut in any graph. Since the maximum multicommodity flow value equals to the value of the minimum fractional multicut, the flow-cut gap is precisely the integrality gap of (LP-multicut).

For undirected graphs, Garg, Vazirani and Yannakakis [45], building on the work of Leighton and Rao [67] and Klein et al. [57] showed that the integrality gap of (LP-multicut) is $O(\log k)$, by providing an efficient LP-rounding algorithm, whose approximation factor is $O(\log k)$. This bound on the integrality gap is almost tight: there is an instance of the minimum multicut problem, for which $\mathsf{OPT} = \Omega(\log k) \cdot \mathsf{OPT}_{LP}$ [67]. The integrality gap of (LP-multicut), and the flow-cut gap for undirected graphs are therefore well understood (to within a constant factor), and stand on $\Theta(\log k)$. The question of whether one can obtain a better than $O(\log k)$-approximation algorithm for undirected multicut by other methods, or perhaps by LP-rounding of a different linear programming relaxation remains wide open. The best currently known hardness of approximation result shows that for some constant c, the problem does not have a c-approximation algorithm, assuming that $\mathbf{P} \neq \mathbf{NP}$ [36]. Under a complexity assumption called the Unique Games Conjecture [55], the undirected multicut problem is hard to approximate to within any constant factor [24, 56]. The status of the Unique Games Conjecture is however still wide open.

The situation is very different in directed graphs. It is easy to obtain a factor k-approximation to the minimum multicut problem, by computing, for each $1 \leq i \leq k$, a minimum s_i–t_i cut E'_i, and returning $\bigcup_{i=1}^{k} E'_i$ as the solution to minimum multicut. Surprisingly, a beautiful construction of Saks et al. [77] shows that this algorithm is close to the best one can achieve via the LP-rounding of (LP-multicut), since the integrality gap of (LP-multicut), and hence the flow-cut gap, can be as large as $k - \epsilon$ for any $\epsilon > 0$ in directed graphs. The number of pairs k in their construction is however quite small when compared to the total number of vertices n in the graph: $k = \Theta(\log n / \log\log n)$, and hence, as a function of n, the lower bound they achieve on the integrality gap is only $\Omega(\log n / \log\log n)$. Unfortunately, [34]

have shown that the integrality gap of (LP-multicut), and therefore the flow-cut gap in directed graphs is at least $\Omega(n^{1/7}/\operatorname{poly}\log n)$[1]. The best current approximation algorithm achieves an $O(n^{11/23} \cdot \operatorname{poly}\log n)$-approximation via LP-rounding of (LP-multicut), thus providing an upper bound of $O(n^{11/23} \cdot \operatorname{poly}\log n)$ on the flow-cut gap [1, 31, 46]. The value of the flow-cut gap for directed graphs therefore remains open, but, unlike undirected graphs, it is polynomially large in n. Minimum multicut in directed graphs is hard to approximate to within factor $2^{\Omega(\log^{1-\epsilon} n)}$ for any constant $\epsilon > 0$, under the plausible complexity assumption that some problems in **NP** do not have efficient randomized algorithms [34].

3. Concurrent flow and sparsest cut

Maximum concurrent flow problem can be seen as multicommodity flow with additional fairness requirements. The input to this problem is the same as in maximum multicommodity flow, but instead of routing maximum amount of flow between all demand pairs, we would like to ensure that **every** demand pair routes a significant amount of flow, and we measure our success by the smallest amount of flow routed between any demand pair. In other words, we would like to maximize a value λ, such that each demand pair (s_i, t_i) can route λ flow units from s_i to t_i simultaneously, and the total flow on any edge e does not exceed its capacity $c(e)$. The linear programming formulation of this problem uses the same notation as in (LP-multi-flow), and is as follows:

$$\begin{array}{lll}
\text{(LP-concurrent-flow)} & \max & \lambda \\
& \text{s.t.} & \\
& \sum_{P\in\mathcal{P}_i} f(P) \geq \lambda & \forall 1 \leq i \leq k \\
& \sum_{P:e\in P} f(P) \leq c(e) & \forall e \in E \\
& f(P) \geq 0 & \forall 1 \leq i \leq k \quad \forall P \in \mathcal{P}_i
\end{array}$$

Often, a more general version of this problem is considered, where each demand pair (s_i, t_i) is associated with a demand value $D_i \geq 0$, and we need to route λD_i flow units from s_i to t_i simultaneously, without violating the edge capacities, for largest possible value λ. Linear program (LP-concurrent-flow) can also be solved efficiently using methods similar to those discussed in Section 1. The dual linear program for (LP-concurrent flow), that we call (LP-spcut), appears below.

$$\begin{array}{lll}
\text{(LP-spcut)} & \min & \sum_{e\in E} c(e)x_e \\
& \text{s.t.} & \\
& \sum_{e\in P} x_e \geq h_i & \forall i : 1 \leq i \leq k, \forall P \in \mathcal{P}_i \qquad (3.1) \\
& \sum_{i=1}^{k} h_i \geq 1 & \qquad (3.2) \\
& x_e \geq 0 & \forall e \in E \\
& h_i \geq 0 & \forall 1 \leq i \leq k
\end{array}$$

This linear program can be seen as a relaxation of a different graph cut problem, called the sparsest cut problem. Suppose we are given an undirected graph $G = (V, E)$ with costs

[1]We say that $f(n) = \operatorname{poly}\log n$ if there is some constant c, such that $f = \Theta((\log n)^c)$.

$c(e)$ on edges $e \in E$, and a collection $\mathcal{M} = \{(s_1, t_1), \ldots, (s_k, t_k)\}$ of demand pairs. For any subset $S \subseteq V$ of vertices, let $\overline{S} = V \setminus S$. Let $E(S, \overline{S})$ denote the set of all edges with exactly one endpoint in S, and let $D(S, \overline{S})$ be the set of all demand pairs (s_i, t_i), where exactly one of s_i, t_i belongs to S. The *sparsity* of S is $\frac{\sum_{e \in E(S,\overline{S})} c(e)}{|D(S,\overline{S})|}$. In the sparsest cut problem, the goal is to find a subset $S \subseteq V$ of vertices of minimum sparsity. If the set $\mathcal{M}$ of the demand pairs contains every pair of vertices of G, then the problem is called *uniform sparsest cut*. The general version of the problem, where $\mathcal{M}$ can be arbitrary, is often called the *non-uniform sparsest cut* problem.

Sparsest cut is one of the central combinatorial optimization problems. It is closely related to the important graph theoretic notions of graph expansion and graph conductance. Approximation algorithms for the sparsest cut problem are often used as subroutines in algorithms for problems arising in many different areas of Computer Science. As an example, one of the most useful paradigms in algorithm design is divide-and-conquer, that often requires a small balanced partition of a given graph G. That is, we need to partition $V(G)$ into two sub-sets V_1, V_2, each of which only contains a constant fraction (say at most $2/3$) of the vertices of G, such that the number of edges $|E(V_1, V_2)|$ is minimized. This problem can be approximately solved by using an approximation algorithm for the sparsest cut problem as a subroutine.

In order to see that (LP-spcut) is a relaxation of the sparsest cut problem, consider any solution S to the sparsest cut problem, and let $E' = E(S, \overline{S})$. For each edge $e \in E$, define a new variable x'_e whose value is 1 if $e \in E'$ and 0 otherwise. For each $i : 1 \leq i \leq k$, define a new variable h'_i, whose value is 1 if $(s_i, t_i) \in D(S, \overline{S})$, and 0 otherwise. Let $D = |D(S, \overline{S})|$. We are now ready to define a solution to (LP-spcut): for each edge $e \in E$, set $x_e = x'_e/D$, and for each $1 \leq i \leq k$, set $h_i = h'_i/D$. It is then easy to see that we have defined a feasible solution to the linear program (LP-spcut), and the value of the solution $\sum_e c(e)x_e = \frac{\sum_{e \in E} c(e)x'_e}{D} = \frac{\sum_{e \in E'} c(e)}{|D(S,\overline{S})|}$ is exactly the sparsity of S. As with undirected multicut, there is an LP-rounding approximation algorithm for the sparsest cut problem, whose approximation factor is $O(\log k)$ in undirected graphs [13, 67, 68], and a matching lower bound of $\Omega(\log k)$ on the integrality gap of (LP-spcut) [67]. Therefore, the flow-cut gap between maximum concurrent flow and sparsest cut in undirected graphs is $\Theta(\log k)$. In a major breakthrough, Arora, Rao and Vazirani [11] designed an $O(\sqrt{\log n})$-approximation algorithm for uniform sparsest cut, by rounding a semidefinite relaxation of the problem. Their algorithm was later generalized to the non-uniform sparsest cut problem, where the approximation ratio becomes $O(\sqrt{\log k} \cdot \log \log k)$ [10]. Somewhat surprisingly, these techniques do not seem to help with the minimum multicut problem, where the best approximation ratio in undirected graphs still stands on $O(\log k)$, and is achieved by an LP-rounding algorithm of [45, 67]. On the negative side, it is known that the sparsest cut problem does not have a factor c-approximation for some specific constant c, unless all problems in **NP** have randomized subexponential time algorithms [2], and this holds even for the uniform sparsest cut problem. Assuming the Unique Games Conjecture, the non-uniform sparsest cut is hard to approximate to within any constant factor [24, 56]. The approximability of the sparsest cut problem remains one of the central open questions in the area of approximation algorithms. Some progress has recently been made on special cases of the problem [9, 48].

For directed graphs, the notion of a sparsest cut can be defined in two distinct ways. In one version of the problem, which we refer to as the *bipartite sparsest cut*, the sparsest cut in a graph is a bipartition of vertices into two sets S and $\bar{S}$ that minimizes the ratio of

$\frac{\sum_{e\in|E(S,\overline{S})|}c(e)}{|D(S,\overline{S})|}$. In the second version, which we refer to as the *non-bipartite sparsest cut*, we need to select a subset E' of edges, minimizing the ratio of $\sum_{e\in E'}c(e)$ to the number of the demand pairs disconnected in $G\setminus E'$. We note that (LP-spcut) is a relaxation of the non-bipartite sparsest cut. In undirected graphs, it is easy to see that the two notions are equivalent, but this is not the case in directed graphs. The best currently known approximation ratio for the non-bipartite sparsest cut is $O(n^{11/23}\operatorname{poly}\log n)$, achieved by LP-rounding of (LP-spcut) [1, 49]. As in minimum multicut, the integrality gap of (LP-spcut) is $\Omega(n^{1/7}/\operatorname{poly}\log n)$ [34]. Therefore, the integrality gap of (LP-spcut), and the flow-cut gap between maximum concurrent flow and sparsest cut in directed graph is polynomial in n. The non-bipartite sparsest cut problem in directed graphs is hard to approximate to within factor $2^{\Omega(\log^{1-\epsilon}n)}$ for any constant $\epsilon>0$, assuming that some problems in **NP** do not have efficient randomized algorithms [34]. The bipartite sparsest cut is known to be hard to approximate to within $2^{\Omega((\log n)^{\epsilon})}$ for some $\epsilon>0$, unless 3SAT has subexponential-time algorithms [23].

4. Integral routing

In this section we consider integral routing problems. We start with the edge-disjoint paths problem, that can be seen as the integral counterpart of maximum multicommodity flow, and discuss several closely related problems, such as node-disjoint paths and congestion minimization. We then consider an integral counterpart of the maximum concurrent flow problem, called integral concurrent flow.

Edge-disjoint paths problem (EDP) is one of the basic problems in integral routing, and we can think of it as an integral version of maximum multicommodity flow. For simplicity, in this section, we assume that all edge capacities are unit. The input to the EDP problem is an n-vertex graph $G=(V,E)$, that can be either directed or undirected, and a collection $\mathcal{M}=\{(s_1,t_1),\ldots,(s_k,t_k)\}$ of k pairs of vertices, that we call *demand pairs*. In order to route a pair (s_i,t_i), we need to select a path P_i connecting s_i to t_i. The goal is to route a maximum possible number of the demand pairs via *edge-disjoint paths*: that is, every edge e may participate in at most one path in the solution. A closely related problem is node-disjoint paths (NDP), defined exactly like EDP, except that the paths chosen to route the demand pairs now need to be *node-disjoint*, so a vertex of G may belong to at most one such path. For directed graphs, the two problems are almost equivalent: an EDP instance $(G,\mathcal{M})$ can be transformed into an instance of the NDP problem, by sub-dividing every edge of G with a new vertex. (This transformation does not preserve the number of vertices, which can grow, so if we are interested in approximation factors as a function of $|V(G)|$, we should be careful when using this transformation). An instance G of NDP in a directed graph can be transformed into an instance of the EDP problem, by replacing every vertex v with a directed edge (a_v,b_v), and every edge (u,v) with an edge (b_u,a_v). For undirected graphs, it is only known that NDP is more general than EDP, as every instance of EDP can be transformed into an instance of NDP via the same transformation, but the transformation in the opposite direction is not known for undirected graphs.

In directed graphs, both NDP and EDP are NP-hard even when the number of the demand pairs is 2 [42]. The following simple algorithm achieves an $O(\sqrt{m})$-approximation for EDP, where m is the number of the graph edges [47, 61–63]. Start with the empty solution. While

at least one demand pair can be routed in G, select a shortest path P, connecting any demand pair (s_i, t_i). Add P to the solution, delete all edges of P from the graph, and delete (s_i, t_i) from the list of the demand pairs that need to be routed. In order to see that this algorithm obtains an $O(\sqrt{m})$-approximation, consider any optimal solution OPT to the problem. In every iteration, if a path P connecting s_i to t_i is added to the solution, then we delete from OPT all paths sharing edges with P, and the path routing the demand pair (s_i, t_i), if such belongs to OPT. As long as P contains fewer than $\sqrt{m}$ edges, we delete at most $\sqrt{m}+1$ paths from OPT in every iteration, while adding at least one path to our solution. Consider now the first iteration where the length of the selected path P is more than $\sqrt{m}$. Since we choose the shortest path routing any demand pair, and all paths that currently belong to OPT can be chosen by the algorithm, every path in OPT contains at least $\sqrt{m}$ edges, and, since these paths are edge-disjoint, OPT contains at most $\sqrt{m}$ paths in total. Therefore, even if we delete all paths from OPT in the current iteration, while adding only one path to the solution, we still preserve the $O(\sqrt{m})$-ratio between the number of paths deleted from OPT and the number of paths added to the solution, thus obtaining an $O(\sqrt{m})$-approximation. Surprisingly, this simple algorithm is almost the best we can hope for: EDP in directed graphs is hard to approximate to within a factor of $\Omega(m^{1/2-\epsilon})$ for any constant ϵ [47]. For the NDP problem, the algorithm described above gives an $O(\sqrt{n})$-approximation, and the problem is hard to approximate in directed graphs to within a factor of $\Omega(n^{1/2-\epsilon})$ for any constant ϵ [47].

While the approximation status of EDP and NDP is well understood in directed graphs, both problems remain wide open in undirected graphs. When the number k of the demand pairs is bounded by a constant, there is an efficient algorithm to solve both NDP and EDP [73, 76]. We discuss this algorithm in more detail in the following section. For general values of k, it is NP-hard to even decide whether all pairs can be simultaneously routed on edge-disjoint paths [51]. The best currently known approximation algorithms achieve an $O(\sqrt{n})$-approximation for both problems [28, 62], while the best current negative result shows that neither problem has an $O(\log^{1/2-\epsilon} n)$-approximation for any constant ϵ, unless all problems in **NP** have randomized algorithms with running time $n^{O(\mathrm{poly}\log n)}$ [4, 5].

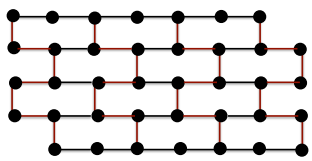
The vertices in the set $T = \{s_1, \ldots, s_k, t_1, \ldots, t_k\}$ are called *terminals*. We will assume for simplicity that all terminals are distinct, that is, $|T| = 2k$, and that each terminal is incident on exactly one edge. This can be assumed without loss of generality, by performing the simple transformation of the input graph G, that preserves the routing solutions: for each terminal $v \in T$, if v participates in z demand pairs, we add z new vertices $v_1, \ldots, v_z$ to G, each of which connects to v with an edge. We then replace v with the vertices $v_1, \ldots, v_z$ in all demand pairs in which v participates, so that each of these new vertices participates in exactly one demand pair. If we add the following integrality constraints to the linear program (LP-multi-flow):

$$f(P) \in \{0, 1\} \qquad \forall 1 \leq i \leq k, \forall P \in \mathcal{P}_i,$$

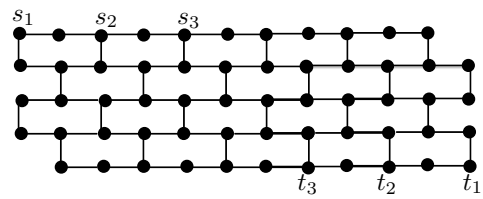
and set the values $c(e)$ for all edges $e \in E$ in the linear program to 1, then we obtain an integral linear program, which is equivalent to EDP. Therefore, (LP-multi-flow) is an LP-relaxation of EDP. The best currently known approximation algorithm for EDP achieves an $O(\sqrt{n})$-approximation by rounding (LP-multi-flow) [28]. Unfortunately, the integrality gap of (LP-multi-flow) is very large even in undirected graphs: if n denotes the number of the graph vertices, and k is the number of the demand pairs, then the integrality gap can be as large as $\Omega(\sqrt{n})$, and as large as $\Omega(k)$ [44]. An example of an instance realizing this integrality gap is a wall graph. Wall graphs play an important role in algorithms for routing

problems and in Graph Minor Theory. They are also among the simplest examples of graphs for which we do not have good approximation algorithms for the EDP problem. A wall of height 5 and width 4 is shown in Figure 4.1(a). A wall of height h and width w can be constructed from a 2-dimensional grid of width $2w$ and height h. Let $C_1, \ldots, C_{2w}$ be the columns of the grid, in their natural left-to-right order. Consider some column C_i, and let $e^i_1, \ldots, e^i_{h-1}$ be the edges of column C_i in their top-to-bottom order. If i is odd, then we delete all edges e^i_j where j is even, and if i is even, we delete all edges e^i_j where j is odd. We also delete all vertices of degree at most 1 in the resulting graph, obtaining a wall of height h and width w. This wall contains h horizontal paths corresponding to the h rows of the grid, that we call the rows of the wall, denoting them by $R_1, \ldots, R_h$ in their natural top-to-bottom order. There are also exactly w disjoint paths connecting the vertices of R_1 to the vertices of R_h, which do not contain the vertices of $R_1 \cup R_h$ as their inner vertices. We call these paths the columns of the wall, and we denote them by $C_1, \ldots, C_w$ in their natural left-to-right order.

In order to define an instance of the edge-disjoint-paths problem, we start with a wall of height $k+2$ and width $2k$. For each $1 \leq i \leq k$, let s_i be the unique vertex in the intersection of R_1 and C_i, and let t_i be the unique vertex in the intersection of R_{k+2} and C_{2k-i+1}. The set of the demand pairs is $\mathcal{M} = \{(s_1, t_1), \ldots, (s_k, t_k)\}$. It is easy to see that the value of the optimal fractional solution for this instance is at least $k/2$: for each $1 \leq i \leq k$, we will define a path P_i connecting s_i to t_i, and we will send $1/2$ flow unit along each such path. We will ensure that every edge of the wall belongs to at most two such paths, obtaining a feasible solution of value $k/2$ to (LP-multi-flow). In order to define path P_i, for $1 \leq i \leq k$, we start from s_i, and follow column C_i, until we reach row R_{i+1}; we then follow R_{i+1} to column C_{2k-i+1}, and column C_{2k-i+1} until we reach t_i. It is immediate to verify that every edge belongs to at most two such paths, and so setting $f(P_i) = 1/2$ for each $1 \leq i \leq k$ gives a feasible solution of value $k/2$ to (LP-multiflow). However, the value of the optimal integral solution is at most 1: assume for contradiction that we can route two demand pairs: (s_i, t_i) and (s_j, t_j), for $i \neq j$, and let P_i, P_j be the two corresponding paths. Let Γ be the cycle that serves as the boundary of the wall. The wall is a planar graph, and has a drawing in the plane, with Γ being the boundary of the outer face - this is the natural drawing, as the one in Figure 4.1(a). The resulting drawings of the paths P_i, P_j have to cross, since their endpoints appear in the circular order (s_i, s_j, t_i, t_j) along Γ. But this is impossible since P_i, P_j are disjoint, and the drawing is planar. Therefore, the integrality gap of (LP-multi-flow) is at least $k/2$, and, since the number of the vertices in our graph is $O(k^2)$, the gap is $\Omega(\sqrt{n})$ as a function of n.



(a) A wall of height 5 and width 4. The columns are shown in red.



(b) An integrality gap example with $k = 3$.

Figure 4.1. A wall graph

Interestingly, even though there are several approximation algorithms achieving constant or polylogarithmic approximation factors for large families of planar graphs, such as grids and grid-like graphs [12, 14, 59, 60], The best currently known approximation ratio for EDP on planar graphs is no better than that for general graphs, namely $O(\sqrt{n})$. Even if the underlying graph is a wall of height $\Theta(\sqrt{n})$ (but the terminals can be located anywhere in the wall and not necessarily on the boundary), no better than $O(\sqrt{n})$-approximation is known, to the best of our knowledge. Closing the gap in our understanding of the approximability of EDP is one of the central problems in the area of graph routing, and a good starting point may be planar graphs or even wall graphs.

The situation with the NDP problem in undirected graphs is very similar: the best current upper and lower bounds on its approximability stand on $O(\sqrt{n})$ and $\Omega(\log^{1/2-\epsilon} n)$ for any constant ϵ, respectively [4, 5, 47, 61–63]. We can again use the multi-commodity flow relaxation of the NDP problem, defined similarly to (LP-multi-flow), except that the capacity constraints are on the vertices and not on the edges of G. This relaxation has an integrality gap of $\Omega(\sqrt{n})$, and the graph realizing this gap is a 2-dimensional $(\sqrt{n} \times \sqrt{n})$-grid. No better than $O(\sqrt{n})$-approximation is known for NDP on planar graphs, and even on grid graphs.

Another important class of graphs is expander graphs. We say that a graph $G = (V, E)$ is an α-expander, iff for any subset $S \subseteq V$ of its vertices with $|S| \leq |V|/2, |E(S, V \setminus S)| \geq \alpha \cdot |S|$. In general, we say that a graph is an expander if it is an α-expander for some fixed constant α independent of the graph size. Both EDP and NDP have polylogarithmic approximation algorithms on bounded-degree expander graphs [20, 21, 43, 58, 67]. Both these problems also have constant-factor approximation algorithms on trees [30, 44], and EDP has constant-factor approximation algorithms on grids and grid-like graphs [12, 14, 59, 60].

Routing with small congestion. Seeing that the status of the EDP problem in undirected graphs is still wide open, it is natural to investigate what happens if we relax the problem requirements slightly, by allowing small *congestion*. We say that a set of paths $\mathcal{P}$ causes edge-congestion c, if every edge belongs to at most c paths in $\mathcal{P}$. When the congestion $c = 1$, we sometimes say that $\mathcal{P}$ causes no congestion. Vertex congestion is defined similarly. It is a common practice to compare a solution to this relaxed version of EDP with an optimal solution that has no congestion. We say that an algorithm achieves an approximation factor α with congestion c for the EDP problem, iff it routes OPT/α demand pairs with congestion c, where OPT is the maximum number of pairs that can be routed with no congestion.

The classical algorithm of Raghavan and Thompson [70] gives a constant factor approximation for EDP with congestion $O(\log n/\log\log n)$. The algorithm performs LP-rounding of (LP-multi-flow), by viewing the values $f(P)$ for each path P as probabilities. Each path $P \in \bigcup_i \mathcal{P}_i$ is selected to the solution independently with probability $f(P)$. If several paths routing the same demand pair are selected, we discard the additional paths arbitrarily. It is not hard to show that with a constant probability we obtain a solution where the number of the demand pairs routed is within a constant of the optimal fractional solution, and each edge participates in at most $O(\log n/\log\log n)$ paths. This randomized rounding scheme can be slightly altered to give, for any congestion value c, a factor $O(cn^{1/(c-1)})$-approximation [15, 16, 62, 80]. More recent result give LP-rounding algorithms for EDP that achieve $O(\mathrm{poly}\log k)$-approximation with smaller congestion [3, 32, 35], with the best current algorithm giving $O(\mathrm{poly}\log k)$-approximation with congestion 2.

An important class of instances of the EDP problem is well-linked instances. We say that

a set T of vertices in graph G is well-linked if for any pair $T_1, T_2 \subseteq T$ of equal-sized subsets of T, there is a set of $|T_1|$ node-disjoint paths connecting the vertices of T_1 to the vertices of T_2 in G. We say that an instance of EDP is well-linked if every terminal participates in exactly one demand pair, and the set of all terminals is well-linked. Chekuri, Khanna and Shepherd [27, 29] have shown an efficient algorithm, that, given any EDP instance $(G, \mathcal{M})$, partitions it into a number of sub-instances $(G_1, \mathcal{M}_1), \ldots, (G_\ell, \mathcal{M}_\ell)$, such that each instance G_i is well-linked, while the sum of the values of the optimal fractional solutions in all these instances is $\Omega(\mathsf{OPT}/\log^2 k)$. Therefore, in order to obtain a polylogarithmic approximation with congestion 2 to EDP, it is enough to find a polylogarithmic approximation with congestion 2 in each such sub-instance separately. The main result of [35] is a structural theorem, that shows that any well-linked instance with k demand pairs contains a large crossbar. The crossbar can be viewed as a degree-3 tree T on $\operatorname{poly}\log k$ vertices, such that every vertex v of T is mapped to a connected subgraph C_v of G, and every edge $e = (u, v)$ of T is mapped to a collection $\mathcal{P}_e$ of $k/\operatorname{poly}\log k$ disjoint paths in G, where each path connects a vertex of C_v to a vertex of C_u. Moreover, each edge of G participates either in at most one graph in $\{C_v\}_{v \in V(T)}$, or in at most one path of $\bigcup_{e \in E(T)} \mathcal{P}_e$, but not both. This crossbar is then exploited to embed an expander X on $k/\operatorname{poly}\log k$ vertices into G with congestion at most 2. Specifically, we select a subset $\mathcal{M}' \subseteq \mathcal{M}$ of $k/\operatorname{poly}\log k$ demand pairs that we will attempt to route. Every vertex v of the expander X is mapped to a connected sub-graph H_v of G, and every edge $e = (u, v)$ of X is mapped to a path P_e in G connecting a vertex of H_v to a vertex of H_u. Each edge of G may participate in up to two sub-graphs H_v, or at most one such sub-graph and at most one path P_e. Each terminal participating in the pairs in $\mathcal{M}'$ belongs a distinct sub-graph H_{v_t} for some $v_t \in V(X)$. The embedding of the expander is performed using the crossbar, building on a beautiful result of Khandekar, Rao and Vazirani [54] on constructing expanders via cut-matching games. Finally, known algorithms for routing on expander graphs are used to find the final routing.

These results demonstrate a fundamental difference between routing with congestion 1 and routing with congestion 2 or higher: Suppose we are given a solution $\mathcal{P}$ to the EDP problem that connects D of the demand pairs with congestion c, and we are interested in obtaining another solution with a lower congestion. By sending $1/c$ flow units along each path in $\mathcal{P}$, we obtain a valid fractional solution to (LP-multi-flow) of value D/c. We can then use the LP-rounding algorithm of [35] to find a solution connecting $\Omega\left(D/(c \operatorname{poly}\log k)\right)$ of the demand pairs with congestion 2. That is, we can lower the congestion to 2 with only a factor $(c \operatorname{poly}\log k)$ loss in the number of the demand pairs routed. However, if we are interested in routing with no congestion, then we may have to lose an $\Omega(\sqrt{n})$-factor in the number of pairs routed, as we can see from the integrality gap example described above: we can view the fractional solution as routing k demand pairs integrally with congestion 2 (by sending 1 flow unit along each path instead of $\frac{1}{2}$), but if we require an integral routing with congestion 1, then at most one pair can be routed.

The $O(\operatorname{poly}\log k)$-approximation algorithm with congestion 2 is close to the best one can hope to obtain from rounding (LP-multi-flow): as discussed above, any sub-polynomial approximation for EDP obtained via this relaxation must incur congestion at least 2. The integrality gap of (LP-multi-flow) is $\Omega\left(\left(\frac{\log n}{(\log\log n)^2}\right)^{1/(c+1)}\right)$ for any constant congestion value c [4], and so the integrality gap for congestion 2 is polylogarithmic. An almost matching hardness of approximation result shows that for any constant ϵ, for any congestion value $c : 1 \leq c \leq O\left(\frac{\log\log n}{\log\log\log n}\right)$, there is no $O\left((\log n)^{\frac{1-\epsilon}{c+1}}\right)$-approximation algorithm for

EDP with congestion c, unless all problems in **NP** have randomized algorithms with running time $(n^{\text{poly}\log n})$ [4]. This gives an $\Omega\left(\log^{(1-\epsilon)/3} n\right)$-hardness of approximation for EDP with congestion 2. These algorithms for EDP were generalized to NDP, giving an $O(\text{poly}\log(k))$-approximation with a constant congestion [26].

Allowing congestion does not seem to help much in directed graphs: EDP remains $n^{\Omega(1/c)}$-hard to approximate even when congestion c is allowed, for any value c between 2 and $\delta \log n/\log\log n$, for some fixed constant δ, unless all problems in **NP** have randomized algorithms with running time $n^{\text{poly}\log n}$ [6, 33], almost matching the $O(cn^{1/(c-1)})$-approximation [15, 16, 62, 80] achievable via the randomized rounding technique.

Congestion minimization Congestion minimization is a natural counterpart of the EDP problem: here, the goal is to route **all** demand pairs, while minimizing the edge congestion. We can slightly alter (LP-multi-flow) to obtain an LP-relaxation for the congestion minimization problem:

$$\begin{array}{lll} \text{(LP-cong-min)} & \min & c \\ & \text{s.t.} & \\ & & \sum_{P\in\mathcal{P}_i} f(P) = 1 \quad \forall 1 \le i \le k \\ & & \sum_{P:e\in P} f(P) \le c \quad \forall e \in E \\ & & f(P) \ge 0 \qquad \forall 1 \le i \le k \quad \forall P \in \mathcal{P}_i \end{array}$$

The randomized rounding algorithm of Raghavan and Thompson [70] gives the best currently known approximation algorithm for the congestion minimization problem, whose approximation factor is $O(\log n/\log\log n)$, by independently choosing, for each $1 \le i \le k$, one path $P \in \mathcal{P}_i$, where path P is chosen with probability $f(P)$. For directed graphs, this algorithm is close to being the best possible, as the problem is known to be hard to approximate to within factor $\Omega(\log n/\log\log n)$ [6, 33]. But for undirected graphs the problem is still wide open, with the best current negative result standing on $\Omega\left(\frac{\log\log n}{\log\log\log n}\right)$-hardness of approximation, unless all problems in **NP** have randomized algorithms with running time $(n^{\text{poly}\log n})$ [7]. Even the integrality gap of (LP-cong-min) for undirected graphs is not well understood: the current upper bound stands on $O(\log n/\log\log n)$, by the algorithm of [70], and the current lower bound is $\Omega\left(\frac{\log\log n}{\log\log\log n}\right)$ [7].

Integral concurrent flow In the integral concurrent flow problem (ICF), we are given an undirected n-vertex graph $G = (V, E)$, a collection $\{(s_1, t_1), \ldots, (s_k, t_k)\}$ of pairs of vertices that we call demand pairs, and a demand value D_i for each $1 \le i \le k$. The goal is to find a maximum value λ, and a collection $\mathcal{P}$ of paths, such that for each demand pair (s_i, t_i) set $\mathcal{P}$ contains at least $\lfloor \lambda \cdot D_i \rfloor$ paths connecting s_i to t_i, and each edge participates in at most one such path. This problem is an integral counterpart of the maximum concurrent flow problem. To the best of our knowledge, no approximation algorithms are known for the problem. As with the EDP problem, we also consider a relaxed version, where a small congestion is allowed on the edges. Chalermsook et al. [22] showed a $\text{poly}\log n$-approximation algorithm for ICF with a constant congestion, by rounding solutions of an LP-relaxation similar to (LP-concurrent-flow). They also showed that for any values η, α, such that $\eta \cdot \alpha \le O(\log\log n/\log\log\log n)$, no efficient algorithm can find an α-approximate

solution with congestion η to ICF unless all problems in **NP** have randomized algorithms with running time $n^{\mathrm{poly}\log n}$.

Chalermsook at al. [22] also consider a more general version of the ICF, called group-ICF, in which, instead of the k pairs of vertices $\{(s_1,t_1),\ldots,(s_k,t_k)\}$, we are given k pairs of vertex subsets, $((S_1,T_1),\ldots,(S_k,T_k))$, so for each $1\leq i\leq k$, $S_i,T_i\subseteq V$. The goal is to find a maximum value λ, and a collection $\mathcal{P}$ of paths, such that for each $1\leq i\leq k$, there are at least $\lfloor\lambda\cdot D_i\rfloor$ paths connecting the vertices of S_i to the vertices of T_i in $\mathcal{P}$, and every edge $e\in E$ belongs to at most one such path. It is easy to see that group-ICF generalizes both the ICF and the EDP problems. We can use an LP-relaxation similar to (LP-concurrent-flow) for the group-ICF problem. When no congestion is allowed, the integrality gap of the relaxation is $\Omega(\sqrt{n})$, even when $k=2$. Moreover, even if we allow congestion c, this ratio can still be as large as $\Omega(n^{1/c+1})$. Chalermsook et al. [22] show that for any $0<\eta\leq O(\log\log n)$ and $\alpha=O\left(n^{1/2^{2\eta+3}}\right)$, no efficient algorithm can find α-approximate solutions with congestion η for group-ICF, unless all problems in **NP** have algorithms with running time $n^{O(\log\log n)}$. Given an optimal integral solution $\mathcal{P}$ to the group-ICF problem instance, let $D=\min_i\{\lfloor\lambda^*\cdot D_i\rfloor\}$ be the minimum number of paths connecting any pair (S_i,T_i) in this solution. Their hardness result only holds for the regime where $D<<k$. They further show that if $D>k\,\mathrm{poly}\log n$, then there is an efficient algorithm that finds a $(\mathrm{poly}\log n)$-approximate solution to group-ICF with constant congestion.

5. Routing with few demand pairs

In this section we consider the NDP problem on undirected graphs when the number k of the demand pairs is bounded by a constant independent of the graph size. (Recall that for directed graphs, NDP is NP-hard even for $k=2$ [42]).

Given a graph G, a *separation* of G is a pair (X,Y) of sub-graphs of G, with $X\cup Y=G$ and $E(X)\cap E(Y)=\emptyset$. The *order* of the separation is $|V(X)\cap V(Y)|$. When the number of the demand pairs is $k=2$, the following beautiful theorem can be used to solve the NDP problem.

Theorem 5.1 ([50, 74, 78, 79, 81]). *Let G be a graph and s_1,t_1,s_2,t_2 four vertices. Assume that there is no separation (X,Y) in G of order at most 3, such that $s_1,t_1,s_2,t_2\in V(X)$ and $X\neq G$. Then either both pairs (s_1,t_1) and (s_2,t_2) can be routed on disjoint paths in G, or there is a drawing of G inside a disc in the plane, with s_1,s_2,t_1,t_2 appearing on the boundary of the disc in this circular order. Moreover, there is an efficient algorithm that either finds the routing or the drawing of G.*

In order to apply the above theorem to the NDP problem instance, we pre-process the input graph G as follows: if there is a separation (X,Y) of G of order at most 3, such that $s_1,t_1,s_2,t_2\in V(X)$ and $X\neq G$, then there must be a separation (X',Y') of G with all the above properties, such that Y' is connected. We delete from G all vertices of $V(Y')\setminus V(X')$, and add all edges connecting every pair of vertices in $V(X')\cap V(Y')$. It is easy to see that the two pairs (s_1,t_1), (s_2,t_2) can be routed on disjoint paths in the new graph iff they can be routed on disjoint paths in the old graph. We repeat this process, until G contains no separation (X,Y) of order at most 3, with $s_1,t_1,s_2,t_2\in V(X)$ and $X\neq G$, and then apply Theorem 5.1 to find the routing.

For the case where $k > 2$, but is still bounded by a constant, Robertson and Seymour [73, 76] have shown an efficient algorithm for NDP, with running time $O(n^3 \cdot f(k))$, where n is the number of graph vertices, and f is some function. This running time was later improved to $O(n^2 \cdot f(k))$ [53]. Before we describe their algorithm, we need to define several graph-theoretic notions.

We start with the notion of treewidth. Intuitively, treewidth measures how close our graph is to a tree: the lower the treewidth value (which is always at least 1), the "closer" our graph is to being a tree. Trees are relatively simple graphs, and many combinatorial optimization problems that are NP-hard on general graphs have efficient algorithms on trees. Many other problems have good approximation algorithms on trees, even if such algorithms are not known for general graphs. Sometimes, when the graphs that we work with are not too complex, techniques used for designing algorithms on trees may still be applicable. It would be therefore useful to have some machinery that allows us to adapt known algorithms for trees to "tree-like" graphs, and to have a formal way to measure the "closeness" of a graph to a tree. The notion of treewidth achieves both these goals: it gives a way to measure the closeness of a graph to a tree, while providing a convenient tree-like representation of the graph, that often allows us to adapt the algorithms known for trees to low-treewidth graphs.

The treewidth of a graph $G = (V, E)$ is typically defined via tree decompositions. A tree-decomposition for G consists of a tree $T = (V(T), E(T))$ and a collection of sets $\{X_v \subseteq V\}_{v \in V(T)}$ called *bags*, such that the following two properties are satisfied: (i) for each edge $(a, b) \in E$, there is some node $v \in V(T)$ with both $a, b \in X_v$ and (ii) for each vertex $a \in V$, the set of all nodes of T whose bags contain a form a non-empty (connected) subtree of T. The *width* of a given tree decomposition is $\max_{v \in V(T)} \{|X_v| - 1\}$, and the treewidth of a graph G, denoted by $\mathrm{tw}(G)$, is the width of a minimum-width tree decomposition for G. There is an interesting connection between graph treewidth and well-linkedness: if w denotes the size of the largest-cardinality well-linked set of vertices in G, then $w \leq tw(G) \leq 4w$.

The problem of computing the treewidth of a graph is NP-hard [8]. When the treewidth value k is bounded by a constant, the treewidth and the corresponding tree decomposition can be computed in time $O(n \cdot f(k))$ for some function f [17, 19, 64, 65, 71, 76]. Using the best currently known approximation algorithms for the vertex version of the sparsest cut problem [40], one can obtain an $O(\sqrt{\log k})$-approximation algorithm for computing treewidth in general graphs, together with the corresponding tree decomposition [18].

Suppose we are given an instance $(G, \mathcal{M})$ of the NDP problem, where $|\mathcal{M}| = k$, and assume that we are given a tree decomposition T of G of width w. Then the problem can be solved in time $O(n) \cdot f(w, k)$ for some function f, via dynamic programming, as follows. Using standard methods, we can transform the tree decomposition T into another tree decomposition T', such that $|V(T')| \leq n$, the width of T' is at most $w + 2k$, every vertex of T' has degree at most 3, and there is one vertex v in T whose degree is 1, and $X_v = \{s_1, \ldots, s_k, t_1, \ldots, t_k\}$. We root the tree T at the vertex v. For each vertex u of T, let S_u be the set of all the vertices of T contained in the sub-tree rooted at u, and let $Y_u = \bigcup_{u' \in S_u} X_{u'}$. We define a graph G_u associated with the vertex u to be the sub-graph of G induced by Y_u. We will think of the vertices of X_u as the terminals for the graph G_u. Since $|X_u| \leq w+2k+1$, there are $2^{O((w+k)\log(w+k))}$ ways to define a matching M between the vertices of X_u (where we allow the matchings to be partial). We say that a matching M is *routable* in G_u iff there is a solution to the NDP problem in graph G_u, where every pair of vertices in M is routed. A *folio* of the vertex $u \in V(T)$, denoted by $\pi(u)$, is the list

of all such matchings M (defined over the set X_u of vertices), such that M is routable in G_u. The main idea of the algorithm is to use dynamic programming in order to compute a folio for every vertex $u \in V(T)$, by processing the tree in the bottom-up fashion. For each matching M in the folio $\pi(u)$, we will also compute and store the set of paths in G_u routing the matching M. Notice that once we compute $\pi(v)$, we can select a matching $M \in \pi(v)$ containing the largest number of the demand pairs from $\mathcal{M}$ to obtain a solution to the NDP problem. For each leaf vertex $u \in V(T)$, the folio $\pi(u)$ can be computed by exhaustively going over all possible matchings M, and for each such matching, checking whether it can be routed in G_u by exhaustive search, since $|V(G_u)| \leq w + 2k$. When a non-leaf vertex u is processed, we need to check all possible ways to combine the matchings in the folios $\pi(u')$, $\pi(u'')$ of the two children u', u'' of u into a single folio $\pi(u)$ of u. Since the sizes of all three folios are bounded by $2^{O((w+k)\log(w+k))}$, and $|X_u| \leq 2k + w$, the running time of this algorithm can be bounded by $f(k + w)$ for some function f, and the overall running time $O(n \cdot f(k + w))$. This gives an efficient algorithm for NDP with constant number of demand pairs on bounded-treewidth graphs. But what about general graphs, whose treewidth may not be bounded by a constant? Robertson and Seymour's Excluded Grid Theorem is a very powerful tool for handling such graphs. The theorem states that there is some function $g : \mathbb{Z}^+ \to \mathbb{Z}^+$, such that for any integer t, every graph of treewidth at least $g(t)$ contains a sub-division of the $(t \times t)$-wall (this is equivalent to saying that G contains a $(t \times t)$-grid as a minor). A long line of work is dedicated to improving the known upper and lower bounds on the function g [25, 38, 39, 52, 66, 72, 75]. The best current bounds show that the theorem holds for $g(t) = O(t^{98} \cdot \text{poly}\log(t))$ [25], and the best negative result shows that $g(t) = \Omega(t^2 \log t)$ must hold [72]. Robertson et al. [72] suggest that this value may be sufficient, and Demaine et al. [37] conjecture that the bound of $g(t) = \Theta(t^3)$ is both necessary and sufficient.

Notice that if the treewidth of G is w, then there is a well-linked set of size $\Omega(w)$ in G. We can then use the machinery developed for approximating EDP and NDP in well-linked instances. The first step in the proof of the excluded grid theorem of [25] constructs a crossbar in G, given the set of $\Omega(w)$ well-linked vertices. This step expands and generalizes the crossbar construction from [26, 35]. In the next step, a new crossbar is constructed, where the underlying tree is a path, and then a result of Leaf and Seymour [66] is used to build a large wall in this new crossbar.

We are now ready to complete the description of the algorithm for NDP when the number k of the demand pairs is bounded by a constant. We use some threshold function $\tau(k)$. If the treewidth of graph G is at most $\tau(k)$, then we run the dynamic programming algorithm described above to solve the NDP problem in time $O(n \cdot f(\tau(k)))$. Otherwise, the treewidth of G is at least $\tau(k)$, and we can find a large wall in G. Using this wall, we can identify an *irrelevant vertex* v in G, such that for any subset $\mathcal{M}' \subseteq \mathcal{M}$ of the demand pairs, the pairs in $\mathcal{M}'$ are simultaneously routable in $G \setminus \{v\}$ iff they are simultaneously routable in G. We then delete the vertex v from graph G and continue. Since the number of iteration is bounded by $|V(G)|$, we will eventually arrive at a graph G whose treewidth is at most $\tau(k)$, and then apply the dynamic programming algorithm to it.

Acknowledgements. Supported in part by NSF grant CCF-1318242. The author thanks Chandra Chekuri and David Kim for their comments on an earlier version of this survey.

References

[1] Amit Agarwal, Noga Alon, and Moses S. Charikar, *Improved approximation for directed cut problems*, STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (New York, NY, USA), ACM, 2007, pp. 671–680.

[2] Christoph Ambuhl, Monaldo Mastrolilli, and Ola Svensson, *Inapproximability results for sparsest cut, optimal linear arrangement, and precedence constrained scheduling*, FOCS '07: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (Washington, DC, USA), IEEE Computer Society, 2007, pp. 329–337.

[3] Matthew Andrews, *Approximation algorithms for the edge-disjoint paths problem via Raecke decompositions*, Proceedings of IEEE FOCS, 2010, pp. 277–286.

[4] Matthew Andrews, Julia Chuzhoy, Venkatesan Guruswami, Sanjeev Khanna, Kunal Talwar, and Lisa Zhang, *Inapproximability of edge-disjoint paths and low congestion routing on undirected graphs*, Combinatorica **30** (2010), no. 5, 485–520.

[5] Matthew Andrews and Lisa Zhang, *Hardness of the undirected edge-disjoint paths problem*, STOC, ACM, 2005, pp. 276–283.

[6] Matthew Andrews and Lisa Zhang, *Logarithmic hardness of the directed congestion minimization problem*, STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing (New York, NY, USA), ACM, 2006, pp. 517–526.

[7] ———, *Hardness of the undirected congestion minimization problem*, SIAM J. Comput. **37** (2007), no. 1, 112–131.

[8] Stefan Arnborg, Derek G Corneil, and Andrzej Proskurowski, *Complexity of finding embeddings in a k-tree*, SIAM Journal on Algebraic Discrete Methods **8** (1987), no. 2, 277–284.

[9] Sanjeev Arora, Rong Ge, and Ali Kemal Sinop, *Towards a better approximation for sparsest cut?*, Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on, IEEE, 2013, pp. 270–279.

[10] Sanjeev Arora, James R. Lee, and Assaf Naor, *Euclidean distortion and the sparsest cut*, STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing (New York, NY, USA), ACM, 2005, pp. 553–562.

[11] Sanjeev Arora, Satish Rao, and Umesh V. Vazirani, *Expander flows, geometric embeddings and graph partitioning*, J. ACM **56** (2009), no. 2.

[12] Yonatan Aumann and Yuval Rabani, *Improved bounds for all optical routing*, Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms (Philadelphia, PA, USA), SODA '95, Society for Industrial and Applied Mathematics, 1995, pp. 567–576.

[13] ———, *An $O(\log k)$ approximate min-cut max-flow theorem and approximation algorithm*, SIAM J. Comput. **27** (1998), no. 1, 291–301.

[14] Baruch Awerbuch, Rainer Gawlick, Tom Leighton, and Yuval Rabani, *On-line admission control and circuit routing for high performance computing and communication*, Proc. 35th IEEE Symp. on Foundations of Computer Science, 1994, pp. 412–423.

[15] Yossi Azar and Oded Regev, *Combinatorial algorithms for the unsplittable flow problem*, Algorithmica **44** (2006), no. 1, 49–66.

[16] Alok Baveja and Aravind Srinivasan, *Approximation algorithms for disjoint paths and related routing and packing problems*, Mathematics of Operations Research **25** (2000), 2000.

[17] Hans L Bodlaender, *A linear time algorithm for finding tree-decompositions of small treewidth*, Proceedings of the twenty-fifth annual ACM symposium on Theory of computing, ACM, 1993, pp. 226–234.

[18] Hans L Bodlaender, John R Gilbert, Hjálmtỳr Hafsteinsson, and Ton Kloks, *Approximating treewidth, pathwidth, and minimum elimination tree height*, Graph-Theoretic Concepts in Computer Science, Springer, 1992, pp. 1–12.

[19] Hans L Bodlaender and Ton Kloks, *Better algorithms for the pathwidth and treewidth of graphs*, Automata, Languages and Programming, Springer, 1991, pp. 544–555.

[20] Andrei Z. Broder, Alan M. Frieze, Stephen Suen, and Eli Upfal, *Optimal construction of edge-disjoint paths in random graphs*, Proc. 5th ACM-SIAM SODA, 1994, pp. 603–612.

[21] Andrei Z. Broder, Alan M. Frieze, and Eli Upfal, *Existence and construction of edge-disjoint paths on expander graphs.*, SIAM J. Comput. (1994), 976–989.

[22] Parinya Chalermsook, Julia Chuzhoy, Alina Ene, and Shi Li, *Approximation algorithms and hardness of integral concurrent flow*, Proceedings of the 44th symposium on Theory of Computing (New York, NY, USA), STOC '12, ACM, 2012, pp. 689–708.

[23] Moses Charikar, Konstantin Makarychev, and Yury Makarychev, *Directed metrics and directed graph partitioning problems*, SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm (New York, NY, USA), ACM, 2006, pp. 51–60.

[24] Shuchi Chawla, Robert Krauthgamer, Ravi Kumar, Yuval Rabani, and D. Sivakumar, *On the hardness of approximating multicut and sparsest-cut*, Comput. Complex. **15** (2006), no. 2, 94–114.

[25] Chandra Chekuri and Julia Chuzhoy, *Polynomial bounds for the grid-minor theorem*, STOC, 2014.

[26] Chandra Chekuri and Alina Ene, *Poly-logarithmic approximation for maximum node disjoint paths with constant congestion*, Proc. of ACM-SIAM SODA, 2013.

[27] Chandra Chekuri, Sanjeev Khanna, and F. Bruce Shepherd, *Multicommodity flow, well-linked terminals, and routing problems*, Proc. of ACM STOC, 2005, pp. 183–192.

[28] Chandra Chekuri, Sanjeev Khanna, and F. Bruce Shepherd, *An $O(\sqrt{n})$ approximation and integrality gap for disjoint paths and unsplittable flow*, Theory of Computing **2** (2006), no. 1, 137–146.

[29] Chandra Chekuri, Sanjeev Khanna, and F. Bruce Shepherd, *The all-or-nothing multicommodity flow problem*, SIAM Journal on Computing **42** (2013), no. 4, 1467–1493.

[30] Chandra Chekuri, Marcelo Mydlarz, and F. Bruce Shepherd, *Multicommodity demand flow in a tree and packing integer programs*, ACM Trans. Algorithms **3** (2007).

[31] Joseph Cheriyan, Howard Karloff, and Yuval Rabani, *Approximating directed multicuts*, Combinatorica **25** (2005), no. 3, 251–269.

[32] Julia Chuzhoy, *Routing in undirected graphs with constant congestion*, Proc. of ACM STOC, 2012, pp. 855–874.

[33] Julia Chuzhoy, Venkatesan Guruswami, Sanjeev Khanna, and Kunal Talwar, *Hardness of routing with congestion in directed graphs*, STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (New York, NY, USA), ACM, 2007, pp. 165–178.

[34] Julia Chuzhoy and Sanjeev Khanna, *Polynomial flow-cut gaps and hardness of directed cut problems*, Journal of the ACM (JACM) **56** (2009), no. 2, 6.

[35] Julia Chuzhoy and Shi Li, *A polylogarithimic approximation algorithm for edge-disjoint paths with congestion 2*, Proc. of IEEE FOCS, 2012.

[36] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis, *The complexity of multiterminal cuts*, SIAM J. Comput. **23** (1994), no. 4, 864–894.

[37] Erik Demaine, MohammadTaghi Hajiaghayi, and Ken-ichi Kawarabayashi, *Algorithmic graph minor theory: Improved grid minor bounds and Wagner's contraction*, Algorithmica **54** (2009), 142–180.

[38] Reinhard Diestel, *Graph theory, 4th edition*, Graduate texts in mathematics, vol. 173, Springer, 2012.

[39] Reinhard Diestel, Tommy R. Jensen, Konstantin Yu. Gorbunov, and Carsten Thomassen, *Highly connected sets and the excluded grid theorem*, J. Comb. Theory, Ser. B **75** (1999), no. 1, 61–73.

[40] U. Feige, M.T. Hajiaghayi, and J.R. Lee, *Improved approximation algorithms for minimum weight vertex separators*, **38** (2008), 629–657.

[41] L.R. Ford and D.R. Fulkerson, *Flows in networks*, Princeton University Press, Princeton, NJ, 1962.

[42] Steven Fortune, John Hopcroft, and James Wyllie, *The directed subgraph homeomorphism problem*, Theoretical Computer Science **10** (1980), no. 2, 111–121.

[43] Alan M. Frieze, *Edge-disjoint paths in expander graphs*, SIAM Journal On Computing **30** (2000), 2001.

[44] Naveen Garg, Vijay V. Vazirani, and Mihalis Yannakakis, *Primal-dual approximation algorithms for integral flow and multicut in trees, with applications to matching and set cover*, ICALP (Andrzej Lingas, Rolf G. Karlsson, and Svante Carlsson, eds.), Lecture Notes in Computer Science, vol. 700, Springer, 1993, pp. 64–75.

[45] ———, *Approximate max-flow min-(multi)cut theorems and their applications*, SIAM Journal on Computing **25** (1996), 698–707.

[46] Anupam Gupta, *Improved results for directed multicut*, SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms (Philadelphia, PA, USA), Society for Industrial and Applied Mathematics, 2003, pp. 454–455.

[47] Venkatesan Guruswami, Sanjeev Khanna, Rajmohan Rajaraman, Bruce Shepherd, and Mihalis Yannakakis, *Near-optimal hardness results and approximation algorithms for edge-disjoint paths and related problems*, Journal of Computer and System Sciences, 1999, p. pages.

[48] Venkatesan Guruswami and Ali Kemal Sinop, *Lasserre hierarchy, higher eigenvalues, and approximation schemes for graph partitioning and quadratic integer programming with PSD objectives*, Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on, IEEE, 2011, pp. 482–491.

[49] Mohammad Taghi Hajiaghayi and Harald Räcke, *An $O(\sqrt{n})$-approximation algorithm for directed sparsest cut*, Inf. Process. Lett. **97** (2006), no. 4, 156–160.

[50] H. A. Jung, *Eine verallgemeinerung des n-fachen zusammenhangs fÃijr graphen*, Math. Ann. **187** (1970), 95—103.

[51] R. Karp, *Reducibility among combinatorial problems*, Complexity of Computer Computations (R. Miller and J. Thatcher, eds.), Plenum Press, 1972, pp. 85–103.

[52] K. Kawarabayashi and Y. Kobayashi, *Linear min-max relation between the treewidth of H-minor-free graphs and its largest grid minor*, Proc. of STACS, 2012.

[53] Ken-ichi Kawarabayashi, Yusuke Kobayashi, and Bruce Reed, *The disjoint paths problem in quadratic time*, Journal of Combinatorial Theory, Series B **102** (2012), no. 2, 424–435.

[54] Rohit Khandekar, Satish Rao, and Umesh Vazirani, *Graph partitioning using single commodity flows*, J. ACM **56** (2009), no. 4, 19:1–19:15.

[55] Subhash Khot, *On the power of unique 2-prover 1-round games*, STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing (New York, NY, USA), ACM, 2002, pp. 767–775.

[56] Subhash A. Khot and Nisheeth K. Vishnoi, *The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into ℓ_1*, FOCS '05: Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (Washington, DC, USA), IEEE Computer Society, 2005, pp. 53–62.

[57] Philip N Klein, Ajit Agrawal, R Ravi, and Satish Rao, *Approximation through multicommodity flow*, FOCS, vol. 31, 1990, pp. 726–737.

[58] Jon Kleinberg and Ronitt Rubinfeld, *Short paths in expander graphs*, In Proceedings of the 37th Annual Symposium on Foundations of Computer Science, 1996, pp. 86–95.

[59] Jon M. Kleinberg and Éva Tardos, *Disjoint paths in densely embedded graphs*, Proceedings of the 36th Annual Symposium on Foundations of Computer Science, 1995, pp. 52–61.

[60] _____, *Approximations for the disjoint paths problem in high-diameter planar networks*, J. Comput. Syst. Sci. **57** (1998), no. 1, 61–73.

[61] Jon Michael Kleinberg, *Approximation algorithms for disjoint paths problems*, Ph.D. thesis, Citeseer, 1996.

[62] Stavros G. Kolliopoulos and Clifford Stein, *Approximating disjoint-path problems using packing integer programs*, Mathematical Programming **99** (2004), 63–87.

[63] Petr Kolman and Christian Scheideler, *Improved bounds for the unsplittable flow problem*, J. Algorithms **61** (2006), no. 1, 20–44.

[64] Jens Lagergren, *Efficient parallel algorithms for tree-decomposition and related problems*, Foundations of Computer Science, 1990. Proceedings., 31st Annual Symposium on, IEEE, 1990, pp. 173–182.

[65] Jens Lagergren and Stefan Arnborg, *Finding minimal forbidden minors using a finite congruence*, Automata, Languages and Programming, Springer, 1991, pp. 532–543.

[66] Alexander Leaf and Paul Seymour, *Treewidth and planar minors*, Manuscript, available at https://web.math.princeton.edu/ pds/papers/treewidth/paper.pdf, 2012.

[67] F. T. Leighton and S. Rao, *Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms*, Journal of the ACM **46** (1999), 787–832.

[68] Nathan Linial, Eran London, and Yuri Rabinovich, *The geometry of graphs and some of its algorithmic applications*, Combinatorica **15** (1995), 215–245.

[69] Karl Menger, *Zur allgemeinen kurventheorie*, Fund. Math **10** (1927), 96–115.

[70] Prabhakar Raghavan and Clark D. Tompson, *Randomized rounding: a technique for provably good algorithms and algorithmic proofs*, Combinatorica **7** (1987), 365–374.

[71] Bruce A Reed, *Finding approximate separators and computing tree width quickly*, Proceedings of the twenty-fourth annual ACM symposium on Theory of computing, ACM, 1992, pp. 221–228.

[72] N Robertson, P Seymour, and R Thomas, *Quickly Excluding a Planar Graph*, Journal of Combinatorial Theory, Series B **62** (1994), no. 2, 323–348.

[73] N. Robertson and P. D. Seymour, *Outline of a disjoint paths algorithm*, Paths, Flows and VLSI-Layout, Springer-Verlag, 1990.

[74] N. Robertson and P.D. Seymour, *Graph minors. IX. Disjoint crossed paths*, J. Comb. Theory Ser. B **49** (1990), no. 1, 40–77.

[75] Neil Robertson and P D Seymour, *Graph minors. V. Excluding a planar graph*, Journal of Combinatorial Theory, Series B **41** (1986), no. 1, 92–114.

[76] Neil Robertson and Paul D Seymour, *Graph minors. XIII. The disjoint paths problem*, Journal of Combinatorial Theory, Series B **63** (1995), no. 1, 65–110.

[77] Michael Saks, Alex Samorodnitsky, and Leonid Zosin, *A lower bound on the integrality gap for minimum multicut in directed networks*, Combinatorica **24** (2004), no. 3, 525–530.

[78] Paul D. Seymour, *Disjoint paths in graphs*, Discrete Mathematics **306** (2006), no. 10-11, 979–991.

[79] Yossi Shiloach, *A polynomial solution to the undirected two paths problem*, J. ACM **27** (1980), no. 3, 445–456.

[80] Aravind Srinivasan, *Improved approximations for edge-disjoint paths, unsplittable flow, and related routing problems*, IEEE Symposium on Foundations of Computer Science, 1997, pp. 416–425.

[81] C. Thomassen, *2-linked graphs*, Erop. J. Combinatorics **1** (1980), 371–378.

Toyota Technological Institute at Chicago, 6045 S. Kenwood Ave., Chicago IL 60637
E-mail: cjulia@ttic.edu

Computing on the edge of chaos: Structure and randomness in encrypted computation

Craig Gentry

Abstract. This survey, aimed mainly at mathematicians rather than practitioners, covers recent developments in homomorphic encryption (computing on encrypted data) and program obfuscation (generating encrypted but functional programs). Current schemes for encrypted computation all use essentially the same "noisy" approach: they encrypt via a noisy encoding of the message, they decrypt using an "approximate" ring homomorphism, and in between they employ techniques to carefully control the noise as computations are performed. This noisy approach uses a delicate balance between structure and randomness: structure that allows correct computation despite the randomness of the encryption, and randomness that maintains privacy against the adversary despite the structure. While the noisy approach "works", we need new techniques and insights, both to improve efficiency and to better understand encrypted computation conceptually.

Mathematics Subject Classification (2010). Primary 68Qxx; Secondary 68P25.

Keywords. Cryptography, complexity theory, homomorphic encryption, software obfuscation, learning with errors (LWE).

1. Introduction

Many results in cryptography are counterintuitive. Alice and Bob can agree on a secret key over a public channel. Alice can prove to Bob that she knows something – say, a proof that P $\neq$ NP – without revealing any details of the proof. Alice can send Bob an encryption of her data $m_1, \ldots, m_t$ such that Bob can compute a succinct encryption of $f(m_1, \ldots, m_t)$ for any function f that he wants, but without Bob learning anything about $m_1, \ldots, m_t$. The last trick is called "fully homomorphic encryption" (FHE). This survey is about FHE and another type of encrypted computation called program obfuscation. Obfuscation allows Alice to encrypt a software program so that the obfuscated program is fully executable but hides essential secrets inside.

Before exploring encrypted computation, let us review some basics about computation and cryptography, illustrated by the story of a young theoretical computer scientist.

1.1. Computation. Young Gauss, the story goes, was challenged by his teacher to add up the numbers from 1 to 100. To his teacher's surprise, Gauss computed the solution almost instantly, while the other pupils toiled for the remainder of the class.

While his classmates added the numbers sequentially, Gauss found a shortcut. He saw that, for even n, the first n numbers can be partitioned into $n/2$ pairs that each add up to

Proceedings of the International Congress of Mathematicians, Seoul, 2014

$n+1$, and that therefore the sum of the first n numbers is $n(n+1)/2$. A mathematician might say that Gauss found a formula or expression for the sum of the first n numbers – namely, $n(n+1)/2$. A computer scientist would add that Gauss also found an *algorithm* or *program*. Moreover, Gauss's algorithm is *efficient*, in contrast to the *inefficient* algorithm used by his classmates.

Gauss's algorithm for adding up the first n numbers takes as input the number n, represented by $k = \log_2 n$ bits (or $\log_{10} n$ decimal digits). The most complex part of Gauss's algorithm is to multiply n and $n+1$, which requires $O(k^2)$ steps using grade-school multiplication. Since the number of computational steps in Gauss's algorithm is only polynomial in the size of the input, we say his algorithm is *polynomial-time*. The sequential algorithm used by his classmates takes at least $n = 2^k$ steps, which is *exponential-time*.

If a problem – such as adding up the numbers 1 to n, or multiplying two numbers – has a polynomial-time algorithm that always solves it, then we say the problem is in the complexity class P (for polynomial-time). BPP, which contains P, is the class of problems solvable by efficient algorithms, which includes probabilistic polynomial-time (PPT) algorithms that may use random coins and only solve the problem with good probability. NP (for "nondeterministic polynomial-time") contains problems that, if you happen to guess the solution, you can verify that it is correct in polynomial time. For example, the integer factorization problem – decomposing an integer N into its prime factors, which is essentially the inverse of multiplication – is in NP, but widely believed not to be in BPP. The biggest open problem in complexity theory is to prove $\mathsf{P} \neq \mathsf{NP}$ (if that is the case).

1.2. Cryptography. Since we have not resolved $\mathsf{P} \stackrel{?}{=} \mathsf{NP}$ and other complexity-theoretic questions, we do not know whether strong cryptography is possible. We might live in any of Impagliazzo's Worlds [23]. Impagliazzo imagined a face-off between Gauss and his teacher in five different worlds, each of which is possible given what we currently know. In "Algorithmica", $\mathsf{P} = \mathsf{NP}$ or some moral equivalent, making much of modern cryptography insecure, and making it virtually impossible for the teacher to stump Gauss. To make the face-off fair, the teacher's problem needs to have a succinct verifiable answer, but any such problem is in NP, hence in P, and therefore is easy for Gauss to solve. At the other extreme, in "Cryptomania", public-key cryptography [12, 33] is possible: two parties can communicate secret messages over public channels. Impagliazzo notes "In Cryptomania, Gauss is utterly humiliated. By means of conversations in class, [the teacher] and his pet student would be able to jointly choose a problem that they would both know the answer to, but which Gauss could not solve." Most cryptographers bet their careers that we live in Cryptomania. But betting against the Gausses of the world is a risky proposition, and so "cryptographers seldom sleep well" [25].

Still, cryptographers soldier on. An early triumph was a paper by Goldwasser and Micali [21] that introduced "probabilistic encryption", defined a rigorous (now standard) notion of security for encryption schemes, and proposed an elegant construction of public-key encryption whose security they provably reduced to a natural, plausible computational assumption: that the quadratic residuosity problem is hard. We review their results here as a vigorous warm-up for recent encrypted computation schemes.

A public-key encryption scheme has three efficient algorithms: a key-generation algorithm K that generates public and secret keys $(\mathrm{pk}, \mathrm{sk})$, an encryption algorithm E that takes pk and a plaintext message m and outputs a ciphertext c, and a decryption algorithm D that takes sk and c and recovers m. It is called "public key", since anyone can use the publicly

available pk to encrypt (without needing any secret knowledge). Of course, for any key pair $(\mathrm{pk}, \mathrm{sk})$ output by K, whenever $c = \mathsf{E}(\mathrm{pk}, m)$, it should hold that $m = \mathsf{D}(\mathrm{sk}, c)$.

Goldwasser and Micali observed that, to be secure, an encryption scheme really should be *probabilistic* – that is, E needs to be randomized, and there must be many ciphertexts for each plaintext. If E were deterministic, an adversary could easily detect whether two ciphertexts encrypt the same thing! To make this intuition more precise, they defined a notion of "semantic security" for encryption in terms of a game between a challenger and an adversary. In the initial phase, the adversary can ask the challenger for encryptions of messages of its choosing. (In the public-key setting, the adversary can generate these encryptions itself.) Then, the adversary generates two equal-length messages m_0, m_1 and asks for an encryption of one of them. The challenger sends a "challenge ciphertext" $\mathsf{E}(m_b)$ for random $b \in \{0, 1\}$, the adversary wins the game if it guesses b, and the scheme is considered semantically secure if the adversary has negligible advantage.

In the Goldwasser-Micali (GM) public-key encryption scheme, Alice samples random prime integers p, q according to an appropriate distribution and sets $N = pq$, samples a uniform $x \in (\mathbb{Z}/N\mathbb{Z})^*$ that is a non-square modulo N but whose Jacobi symbol $\left(\frac{x}{N}\right)$ equals 1, and publishes (N, x) as her public key. Bob encrypts $m \in \{0, 1\}$ for Alice by sampling random $r \in (\mathbb{Z}/N\mathbb{Z})^*$ and sending the ciphertext $c \leftarrow x^m \cdot r^2 \in (\mathbb{Z}/N\mathbb{Z})^*$. That is, an encryption of 0 is a square, and an encryption of 1 is a non-square (with Jacobi symbol 1). Alice decrypts to recover m by distinguishing whether c is a square modulo the secret prime factor p (e.g., by using Gauss's quadratic reciprocity theorem).

The quadratic residuosity problem is related to the integer factorization problem. The problem is: given a composite integer $N = pq$ (but not the prime factors p and q) and an element $x \in (\mathbb{Z}/N\mathbb{Z})^*$ whose Jacobi symbol is 1 (where N and x are sampled according to appropriate distributions), decide whether x is a square in $(\mathbb{Z}/N\mathbb{Z})^*$. The quadratic residuosity assumption is that the quadratic residuosity problem is hard (not in BPP). To put it another way, the assumption is that, against all PPT adversaries, the subset of squares modulo $N = pq$ is pseudorandom among the set of elements with Jacobi symbol 1. The assumption is clearly stronger than factoring, but it seems like a safe assumption, since we do not know an actual algorithm to solve it that is significantly faster than factoring. For us, the assumption has the added appeal of taunting our adversary Gauss, since he can use his quadratic reciprocity theorem to compute the Jabobi symbol of x modulo N without knowing N's factorization, but this does not help him since we always fix $\left(\frac{x}{N}\right) = 1$.

To reduce the semantic security of their scheme to quadratic residuosity, Goldwasser and Micali use a "hybrid argument" approach that has become standard. Assume that our adversary Gauss can break the cryptosystem – i.e., can distinguish encryptions of 0 from encryptions of 1. Consider two different games, Game 0 and Game 1. In Game 0, we generate the public key (N, x) and a challenge ciphertext (an encryption of a random bit $m \in \{0, 1\}$) for Gauss in the correct way. By assumption, Gauss should be able to guess m with noticeable advantage. In Game 1, however, we generate the public key (N, x) in a different way. Specifically, we make x a square in $(\mathbb{Z}/N\mathbb{Z})^*$, and generate the challenge ciphertext by encrypting m using the normal encryption procedure, as if (N, x) were a valid public key. In Game 1, encryptions of 0 and encryptions of 1 have the same distribution (either way, the ciphertext is a random square), and thus Gauss cannot have any advantage guessing m. Thus, Gauss's success probability noticeably differs in Games 0 and 1. To construct a PPT algorithm to decide whether x is a non-square or square (i.e., whether we are in Game 0 or Game 1), we simply use Gauss's performance to help us distinguish. This

bases the security of GM on quadratic residuosity.

1.3. Homomorphic encryption. The GM scheme has a curious bonus feature: it is *malleable*. It allows anyone to manipulate (in limited but meaningful ways) what is encrypted, even without knowing the secret key: to *compute on encrypted data*. Specifically, suppose that c_1 is a GM encryption of $m_1 \in \{0,1\}$, and c_2 is a GM encryption of m_2 – that is, $c_1 = x^{m_1} \cdot r_1^2$ and $c_2 = x^{m_2} \cdot r_2^2$ for some $r_1, r_2 \in (\mathbb{Z}/N\mathbb{Z})^*$. We can increment the plaintext by multiplying the ciphertext by x, without even knowing what the plaintext is. The new ciphertext $c \leftarrow c_1 \cdot x = x^{m_1+1} \cdot r_1^2$ encrypts $m_1 + 1$. Also, we can add plaintexts by multiplying the corresponding ciphertexts: $c \leftarrow c_1 \cdot c_2 = x^{m_1+m_2} \cdot (r_1 r_2)^2$ encrypts $m_1 + m_2$. These plaintext additions are in $\mathbb{Z}/2\mathbb{Z}$, since x^2 is an encryption of 0. Interestingly, GM allows an unlimited number of plaintext additions, but GM's overall malleability is limited. GM can compute linear functions on encrypted data, but it does not (for example) provide any way to operate on two ciphertexts so as to multiply the two plaintexts.

Rivest, Adleman and Dertouzos [32] saw the potential of computing on encrypted data a few years earlier in 1978, shortly after the invention of the RSA public-key encryption scheme [33], which allows multiplications of plaintexts but not additions. They wondered whether it could be possible to construct an encryption scheme that is *completely* malleable, that allows *unlimited* computations on encrypted data. They called such a scheme a "privacy homomorphism". These days, we call it "fully homomorphic encryption" (FHE), where "fully" means it allows any computation over encrypted values. (GM is "additively homomorphic" and RSA is "multiplicatively homomorphic".) They also foresaw that an FHE scheme would have amazing applications. It took more than 30 years after Rivest et al. proposed the notion to discover the first plausible FHE scheme [16]. Now that we have discovered plausible constructions, we have made tremendous progress improving them, but still have far to go.

Before we address what FHE can do, let us be more precise about what it is. In this survey, an FHE scheme is first of all a public-key encryption scheme with the usual algorithms K, E, and D. Let $\mathcal{M}$ and $\mathcal{C}$ be the message space and ciphertext space of the scheme. Let us say that a ciphertext $c \in \mathcal{C}$ encrypts a message $m \in \mathcal{M}$ under key $(\mathrm{pk}, \mathrm{sk})$ if decryption returns $m \leftarrow \mathsf{D}(\mathrm{sk}, c)$. The special feature of an FHE scheme is that it comes equipped with a *fourth* efficient algorithm, called *Evaluate* and denoted by V, such that for any valid key pair $(\mathrm{pk}, \mathrm{sk})$, any t (for any t) encryptions $c_1, \ldots, c_t$ of any messages $m_1, \ldots, m_t \in \mathcal{M}$ under $(\mathrm{pk}, \mathrm{sk})$, and for any t-ary function $f : \mathcal{M}^t \to \mathcal{M}$, $\mathsf{V}(\mathrm{pk}, f, c_1, \ldots, c_t)$ outputs a ciphertext c that encrypts $f(m_1, \ldots, m_t)$. Crucially, Evaluate is a public algorithm that anyone can execute without the secret key, and of course we want the encryption scheme to be semantically secure despite its availability. In short, an FHE scheme allows computation of any function f *inside* an "impenetrable box" of encryption.

We can describe FHE in terms of a commutative diagram.

$$\begin{array}{ccc} \mathcal{C}^t & \xrightarrow{\mathsf{V}(\mathrm{pk}, f, \cdot, \ldots, \cdot)} & \mathcal{C} \\ \Big\downarrow{\scriptstyle \mathsf{D}(\mathrm{sk}, \cdot, \ldots, \cdot)} & & \Big\downarrow{\scriptstyle \mathsf{D}(\mathrm{sk}, \cdot)} \\ \mathcal{M}^t & \xrightarrow{f(\cdot, \ldots, \cdot)} & \mathcal{M} \end{array}$$

The diagram is meant to convey that, for any key, messages, ciphertexts, and function f, the order of decryption and applying f does not matter: either way we end up with

$f(m_1, \ldots, m_t)$. An analogous commutative diagram with encryption instead of decryption does not work. Although it is true that the order of encryption and applying f does not matter in the sense that (either way) we end up with an encryption of $f(m_1, \ldots, m_t)$, the actual ciphertexts might be different. (Recall that having many different ciphertexts for each message is essential for an encryption scheme to be semantically secure.)

Later in the survey, we will see in detail how to construct an FHE scheme. At this point, we must keep the reader in suspense.

1.3.1. Applications of homomorphic encryption. An exciting potential application of FHE is preserving privacy online, which is more relevant now than ever before. For example, we seem to heading toward widespread acceptance of cloud computing, where users put their data online "in the cloud" for convenience and availability. Putting everything online unencrypted is to risk an Orwellian future, not just because the corporation hosting our data may misuse it, but also because a government may strong-arm the corporation into providing a backdoor. For certain types of data, such as medical records, storing them off-site unencrypted may be illegal. On the other hand, encrypting one's data seems to nullify the benefits of the "computing" part of cloud computing. Unless I give the cloud my secret decryption key (sacrificing my privacy), how can I expect the cloud to do any meaningful processing of my encrypted data? Fully homomorphic encryption provides a way out of this false dilemma. If I want to make some query f on my encrypted data, I can just send a description of f to the cloud, which uses the Evaluate algorithm to derive an encryption of $f(m_1, \ldots, m_t)$, which is the response to my query.

In addition to encrypting my data, I can encrypt my query f (under the same pk). More broadly, I can encrypt a *program* P, so that the cloud can execute P on unencrypted data or data encrypted under the same pk, and output the encrypted result. At first, this fact may seem surprising, but it is just an application of Turing's idea that a program can be viewed just another type of data to be processed by a universal Turing machine. (In more modern terms, a program can be read and executed by an interpreter program.)

The applications of FHE may seem counterintuitive and hard to believe. In a world with FHE – call it "Cryptomegalomania" – cryptography flexes its muscles and sticks its tongue out at young Gauss. Gauss might have the last laugh though. Current FHE schemes are too impractical to realize all of the applications that are possible in principle. Developing a significantly faster FHE scheme is an interesting mathematical problem that also has high stakes for society.

1.3.2. Shortcomings of homomorphic encryption. Besides high overhead, there are two related "problems" with FHE.

The first problem is that Evaluate always has an *encrypted output*. This is, in some sense, optimal for security: nothing is ever revealed to anyone but the secret key holder. But it is often sub-optimal for functionality. Sometimes it is useful to reveal some (carefully controlled) *unencrypted* information to the Evaluator. This especially true for *encrypted programs*. One might like to hide (encrypt) certain aspects of a program (e.g., to prevent it from being semantically deconstructed) while preserving its functionality as a fully executable program with unencrypted inputs and outputs.

The second problem is that, while FHE can handle general computations "efficiently" in the sense of "polynomial-time", FHE cannot exploit certain optimizations essential to the practicality of computation in modern computing environments. Specifically, FHE needs to

put a function f or program P into a special format – called a boolean or arithmetic *circuit* – before it can be processed.[1] In a circuit evaluation of f, the number of computational steps does not depend on the input x. For the security of FHE, this is necessary: if the run time of Evaluate depended on the particular value of (encrypted) x, it would reveal something about x. However, it also means that Evaluate's run time depends on the *worst-case* x's; Evaluate can never take a shortcut for "easy" inputs. Similarly, FHE cannot do *random access* (as in a random access machine (RAM)) over encrypted data, since FHE does not allow the Evaluator to learn unencrypted data-dependent addresses. Nor does FHE allow an Evaluator to exploit an *inverted index*, which helps make searches (like web searches) over huge data-sets practical.

1.4. Program obfuscation. Using FHE, we can generate encrypted programs that have *encrypted output*. But is there some way to generate encrypted programs that have *unencrypted output*? To put it another way: Is there any meaningful sense in which we can "encrypt" a program while preserving its functionality (input/output behavior) as a fully executable program? This is the seemingly-paradoxical and hard-to-define goal of *program obfuscation*.

Program obfuscation may sound impossible to achieve, and indeed some notions of obfuscation are. For example, consider a program P that prints its own code. Since any obfuscation $O(P)$ of P must have the same functionality as P, $O(P)$ reveals P completely. Barak et al. [4] showed that some programs are unobfuscatable even without being so exhibitionist. They showed that, assuming one-way functions (functions that are hard to invert), there are *unlearnable* programs P (programs for which no PPT algorithm can recover P or any code equivalent to P just from oracle access to P) that can be completely recovered from any code that implements them. Obfuscation is impossible in an "absolute" sense: for some programs, any obfuscation reveals everything.

However, it turns out that obfuscation is possible in a "relative" sense. To understand this notion of encrypting a program, let us revisit what it means to encrypt a message. Goldwasser and Micali called an encryption scheme "semantically secure" if a PPT adversary has negligible advantage of winning the following game: the adversary picks two equal-length messages m_0, m_1, the challenger encrypts one of them, and the adversary tries to guess which one. They need the "equal-length" message requirement, because a ciphertext always reveals some information about the message it encrypts – namely, an upper bound on its length. Similarly, an obfuscated program always reveals something about the original program – an upper bound on its size, and also the program's input/output behavior. Accordingly, Barak et al. [4] defined an analogue of semantic security for programs via a similar game: the adversary picks two equal-size functionally-equivalent programs (represented as circuits C_0, C_1), the challenger obfuscates one of them, and the adversary tries to guess which one. The obfuscator is considered secure if every PPT adversary has negligible advantage of winning the game. This notion is called *indistinguishability obfuscation* (IO).

It is not obvious that IO is actually useful. An IO obfuscator does not guarantee it will hide any secrets residing in the program. It does not provide any absolute guarantees about the quality of the obfuscation. However, IO provides a strong relative guarantee – namely, an indistinguishability obfuscator is a "best-possible" obfuscator: it is as good as any other obfuscator of roughly the same complexity [4, 22]. To see this, suppose O is a secure indistinguishability obfuscator. Suppose $BO(\cdot)$ is the actual best obfuscator of a

[1] We will discuss circuits in more detail in Section 2.

certain complexity of circuits of a certain size, whereas $Pad(\cdot)$ merely increases the size of circuits the same amount as $BO(\cdot)$. Then, for any circuit C, the circuits $BO(C)$ and $Pad(C)$ are the same size and have the same functionality, and so $O(BO(C))$ and $O(Pad(C))$ are indistinguishable. Since they are indistinguishable, $O(Pad(C))$ obfuscates C as well as $O(BO(C))$, which obfuscates C as well as $BO(C)$.

Although IO only provides a relative guarantee of security, it can be used to construct schemes having absolute guarantees. For example, Garg et al. [15] showed how to use IO to construct a functional encryption scheme [5]: a public-key scheme administered by an authority that chooses a function f and distributes secret keys to users such that a user with sk_y associated to string y can recover exactly $f(x, y)$ from a ciphertext encrypting x. For example, y might specify a user's security clearance, and f might specify a redaction policy, such that user y obtains only the portion of document x for which it has clearance. Obfuscation can also be used to "fix" some of the problems with FHE. For example, it can be used to allow encrypted computation in the RAM model of computation (rather than circuits) [1, 18].

Garg et al. [15] recently found the first plausibly secure construction of IO. Here is a very brief overview of how their scheme works. First, they show how to "bootstrap" IO for NC^1 (logarithmic depth) circuits to IO for general circuits. Specifically, the obfuscation of a circuit C consists of encryptions of C under two FHE key pairs $(\mathrm{sk}_0, \mathrm{pk}_0)$, $(\mathrm{sk}_1, \mathrm{pk}_1)$ and an obfuscated conditional decryption circuit $O(\mathsf{ConD})$ (to be described momentarily). The Evaluator computes the encrypted program C on its input under both FHE public keys, and feeds the resulting ciphertexts, with a "proof" that they were computed correctly, as input to $O(\mathsf{ConD})$, which decrypts one ciphertext using sk_0 if the proof verifies. Garg et al. use the fact that ConD can be implemented in NC^1 for known FHE schemes. Assuming O is a secure IO for NC^1, they show that a PPT attacker cannot distinguish whether the FHE secret key inside $O(\mathsf{ConD})$ is sk_0 or sk_1, since either way ConD's output is the same. This shell game shows that sk_0 is hidden, and forms part of their hybrid security proof for IO for general circuits.

Next, Garg et al. present an indistinguishability obfuscator for NC^1 circuits. Their NC^1 obfuscator uses a graded encoding scheme by Garg et al. [14]. A graded encoding scheme is similar to a homomorphic encryption scheme, with the important difference that it comes equipped with zero test that allows anyone to efficiently distinguish when the encoded value is 0. This zero test allows some unencrypted information to leak (unlike FHE), but schemes using graded encodings are carefully designed to ensure that (hopefully) this leakage can only occur when the Evaluator computes over the encodings in a permitted way. Currently, known schemes for IO for NC^1 have security based on unconventional assumptions about graded encodings.

Since Garg et al.'s obfuscation construction, there have been some improvements both in security and efficiency, but both aspects are still worse than for FHE, in part because current obfuscation schemes use FHE as a *component*. This is a young and active area of research.

1.5. "Computing on the Edge of Chaos" and "Structure and Randomness". We now begin turning to the construction of FHE and obfuscation schemes. Before we begin in earnest, let us start with a high-level intuition for how current FHE schemes (and the obfuscation schemes derived from them) work. Current FHE schemes all use essentially the same "noisy" approach. They encrypt via a noisy encoding of the message: by sending the message to a ciphertext that is similar to a perturbed codeword in an error-correcting code. The

decrypter recovers the message by recovering the noise. The public key is, in some sense, a "bad" basis of the error-correcting code, which permits efficient encryption but does not permit efficient correction of errors. By careful manipulation of the ciphertexts, an Evaluator can add and multiply the underlying plaintexts while increasing the noise by only a small amount. Furthermore, when the noise becomes almost large enough to drown out the signal (the message), the Evaluator can apply an operation called "bootstrapping" to "refresh" the noisy ciphertext: to generate a new ciphertext that encrypts the same message but with less noise. In short, the "noise" turns out to be both a boon and a bane. The noise hides the message from adversaries. However the noise lies behind the impracticality of current schemes: it makes ciphertexts large and it requires computationally expensive steps to bound the noise as computations are performed.

The phrases "computing on the edge of chaos" and "structure and randomness" capture some intuitions that I have about encrypted computation, and the possibility that a noisy approach may be necessary. Of course, these intuitions may be illusions. I would like nothing more than for someone to find a radically different way of constructing fully homomorphic encryption and obfuscation schemes that escapes the current paradigm of using noisy, approximate homomorphisms. Consider the title of this paper a provocation, a challenge.

My (not very strongly held) intuition, for what it's worth, is that "exact" mathematical structures – e.g., exact rather than approximate homomorphisms of the kind used in previous weakly homomorphic encryption schemes such as Goldwasser-Micali – seem either too rigid (e.g., they allow only additive but not multiplicative homomorphism) or too permissive (e.g., they allow full homomorphism but enable trivial linear algebra attacks). Instead, for robust encrypted computation, we seem to need mathematical structures that can be inexact without simply being wrong – that is, structures that noisily remain close to exact solutions.

To be secure under Goldwasser and Micali's notion of "semantic security", an encryption scheme must be probabilistic – i.e., it must use randomness in encryption. But getting this randomness to play nicely with the structure we need for correct computation is a delicate balance, and it raises certain questions: What happens to the randomness when we do homomorphic operations on ciphertexts? Does the randomness mix with the structured part of the ciphertexts, or does it somehow remain cordoned off? If the former, how is the structure preserved (so as to allow correct decryption)? If the latter, how does the randomness remain safely cordoned off despite performing complex general computations? (It seems like general computation would induce a lot of mixing.) Also, in the latter case, how does the scheme remain secure – for example, how does it remain secure against linear algebra attacks if the randomness is perpetually isolated to certain coordinates? In the noisy approach to homomorphic encryption, the randomness indeed mixes with the structure (in particular, with the message), but the randomness is always kept small so that it does not overwhelm the structure.

I thought "computing on the edge of chaos" would be a fun and original way to describe the current approach to encrypted computation, but it turns out the phrase has already been taken. Apparently, it refers to a critical phase transition point in cellular automata between overly ordered and completely chaotic where the automata become capable of universal computation, and more broadly refers to the notion that dynamic "lifelike" systems, such as the economy or human brain, are healthiest when they are "poised on the edge of chaos". The notion seems intuitively appealing, though there has been pushback against it as being unrigorous and unsubstantiated. The idea that the noisy approach to encrypted computation somehow exploits a phase transition between order and chaos also seems intuitively

appealing, if even more unsubstantiated.

1.6. Roadmap. In the rest of the survey, we will limit our focus to FHE. We will describe in depth how to construct an FHE scheme with security provably based on the hardness of the so-called learning with errors (LWE) problem.

2. Circuits and homomorphic encryption

We touched upon circuits and homomorphic encryption in the Introduction. Here, we discuss them more formally.

2.1. Circuits. Before we can specify how to Evaluate a function using homomorphic encryption, we need to be more explicit about our *model of computation*. The canonical theoretical representation of a computer is the Turing machine, described by Alan Turing in the 1930's. It handles general computations, and is as efficient as modern random access memory (RAM) computers up to polynomial factors (assuming the RAM computer's memory is not pre-loaded). However, in this survey, we will primarily use a mathematically cleaner representation of algorithms, called a boolean or arithmetic *circuit*. Circuits also handle general computations, and almost as efficiently as Turing machines. In particular, if there is a Turing machine program that always evaluates a function f in at most T_f steps, then there is a circuit for f that has size $O(T_f \cdot \log T_f)$ [30].

An arithmetic circuit is a remarkably simple and mathematically clean way of representing a program. It is typically just a composition of addition gates (which take several inputs and output their sum), multiplication gates (which take several inputs and output their product), and scalar multiplication gates (which take one input and multiply it by a scalar), where these operations are performed over some ring. The gates are typically arranged into levels, so that the outputs of gates at level i are inputs to gates at level $i + 1$ unless i is the last level of the circuit. The circuit cannot contain any loops (it is a directed acyclic graph), but one can reuse the output of a gate as input to multiple higher-level gates. The number of gates is called the *size* of the circuit, and the number of levels is called the *depth*. Notice that, since the circuit just uses addition and multiplication, the output of each gate has a nice mathematical interpretation: it is simply a multivariate polynomial (evaluated at the inputs).

When the ring is $\mathbb{F}_2$ and each gate has at most two inputs, we call the circuit a boolean circuit. Interestingly, any boolean function can be computed using a circuit composed entirely of NAND gates. For $x, y \in \{0, 1\}$, $\text{NAND}(x, y) = 1 - x \cdot y \in \{0, 1\}$. Restricting to $\{0, 1\}$, we can implement NAND over any ring.

It may be surprising that multivariate polynomials representable by polynomial-size circuits, even boolean circuits of NAND gates, are adequate to represent polynomial-time computation.[2] However, a multivariate polynomial with low circuit complexity may be very complex by other measures. Even when the circuit has polynomial size, the multivariate polynomials it represents may have an exponential number of monomials. Moreover, over large fields, the degree of the polynomials may be exponential in the depth of the circuit,

[2]Leslie Lamport, in his essay *How to Tell a Program from an Automobile*, remarked that "An automobile runs, a program does not. (Computers run, but I'm not discussing them.) ... An automobile is a piece of machinery, a program is some kind of mathematical expression". Lamport's observation becomes especially clear when the program is represented as a circuit, which in turn represents nothing more than a set of multivariate polynomials.

since each level of multiplication gates may double the degree.

2.2. Homomorphic encryption. A homomorphic encryption scheme is a tuple of four probabilistic polynomial time (PPT) algorithms $(\mathsf{K}, \mathsf{E}, \mathsf{D}, \mathsf{V})$. In this survey, the message space $\mathcal{M}$ of the scheme will always be some ring and our computational model will be arithmetic circuits over this ring (e.g., addition, multiplication and NAND gates).

- $\mathsf{HE.K}$ takes the security parameter λ (and possibly other parameters of the scheme) and produces a secret key sk and a public key pk.
- $\mathsf{HE.E}$ takes pk a message $m \in \mathcal{M}$ and produces a ciphertext c which is the encryption of m.
- $\mathsf{HE.D}$ takes sk and a ciphertext c and produces a message m.
- $\mathsf{HE.V}$ takes pk, an arithmetic circuit f over $\mathcal{M}$, and ciphertexts $c_1, \ldots, c_t$, where t is the number of inputs to f, and outputs a ciphertext c.

Roughly speaking, the security parameter λ specifies the security level of the scheme. The algorithms of the scheme should take time $\mathrm{poly}(\lambda)$, but any known algorithms to attack the scheme should take time super-polynomial in λ, preferably exponential (say 2^{λ}) time.

Definition 2.1 (Correctness and Compactness). We say that a homomorphic encryption scheme $(\mathsf{K}, \mathsf{E}, \mathsf{D}, \mathsf{V})$ *correctly evaluates* a circuit family $\mathcal{F}$ if for all $f \in \mathcal{F}$ and for all $m_1, \ldots, m_t \in \mathcal{M}$ it holds that if sk, pk were properly generated by K with security parameter λ, and if $c_i = \mathsf{E}(\mathrm{pk}, m_i)$ for all i, and $c = \mathsf{V}(\mathrm{pk}, f, c_1, \ldots, c_t)$, then

$$\Pr[\mathsf{D}(\mathrm{sk}, c) \neq f(m_1, \ldots, m_t)] = \mathsf{negl}(\lambda) \ ,$$

where the probability is taken over all the randomness in the experiment.

We say that the scheme *compactly evaluates* the family if in addition the run time of the decryption circuit only depends on λ and not on its input.

The notation $\mathsf{negl}(\lambda)$ means the function grows more slowly than the inverse of any polynomial: $\mathsf{negl}(\lambda) = O(1/\lambda^c)$ for any constant c.

The reason for the compactness requirement is that homomorphic encryption is uninteresting without it. If the ciphertext size could depend on the circuit size, we could just set $c = (f, c_1, \ldots, c_t)$, and decrypt c by decrypting the c_i's and applying f. Obviously such a scheme is useless for delegation of computation, since the decrypter rather than the Evaluator performs all of the computation.

Much of this survey will focus on the construction of a *leveled* fully homomorphic scheme, where the parameters of the scheme depend (polynomially) on the depth (but not the size) of the circuits that the scheme is capable of evaluating.

Definition 2.2 (Leveled FHE). We say that a family of homomorphic encryption schemes $\{\mathcal{E}^{(L)} : L \in \mathbb{Z}^+\}$ is leveled fully homomorphic if, for all $L \in \mathbb{Z}^+$, they all use the same decryption circuit, $\mathcal{E}^{(L)}$ compactly evaluates all circuits of depth at most L, and the computational complexity of $\mathcal{E}^{(L)}$'s algorithms is polynomial (the same polynomial for all L) in the security parameter, L, and (in the case of the evaluation algorithm) the size of the circuit.

In a "pure" FHE scheme, the complexity of the algorithms (except for Evaluate) is independent of L.

We use Goldwasser and Micali's notion of semantic security [21].

Definition 2.3. A homomorphic scheme is secure if any PPT adversary that first gets a properly generated pk, then specifies $m_0, m_1 \in \mathcal{M}$ and finally gets $\mathsf{E}(\mathsf{pk}, m_b)$ for random b, cannot guess b with probability $> 1/2 + \mathsf{negl}(\lambda)$.

Of course, the adversary can try to use the additional Evaluate algorithm to win the semantic security game.

3. Learning with Errors (LWE)

As we saw in the Introduction, when cryptographers construct an encryption scheme, they try to *prove* that the scheme is secure as long as a natural problem (such as quadratic residuosity) is hard to solve. This proof is called a *reduction*. Here, we describe a natural problem called *learning with errors* (LWE). Later, we will show how to construct public-key and homomorphic encryption schemes whose security reduces to it. We also review some evidence that LWE is a hard problem.

The LWE problem was introduced by Regev [31]. Informally, the "search" version of LWE is about solving "noisy" systems of linear equations. The problem is to recover a n-dimensional vector $\vec{s}$ over $\mathbb{Z}/q\mathbb{Z}$ from many pairs $(\vec{a}_i, b_i)$, where the $\vec{a}_i$'s are sampled as uniformly random vectors over $\mathbb{Z}/q\mathbb{Z}$, and b_i is set to $\langle \vec{a}_i, \vec{s} \rangle + e_i \in \mathbb{Z}/q\mathbb{Z}$ for some "error" e_i of small magnitude ($\ll q$). If not for the errors, we could recover $\vec{s}$ efficiently using Gaussian elimination after receiving about n equations. Introducing error seems to make the problem hard.

More formally, LWE is typically defined as a "decision" problem as follows.

Definition 3.1 (LWE). For security parameter λ, let $n = n(\lambda)$ be an integer dimension, $q = q(\lambda) \geq 2$ be an integer, and $\chi = \chi(\lambda)$ be a distribution over $\mathbb{Z}$. The $\mathrm{LWE}_{n,q,\chi}$ problem is to distinguish the following two distributions:

(1) Output $(\vec{a}_i, b_i)$ sampled uniformly from $(\mathbb{Z}/q\mathbb{Z})^{n+1}$.

(2) For fixed uniform $\vec{s} \leftarrow (\mathbb{Z}/q\mathbb{Z})^n$, sample $\vec{a}_i \leftarrow (\mathbb{Z}/q\mathbb{Z})^n$ uniformly, sample $e_i \leftarrow \chi$, set $b_i = \langle \vec{a}_i, \vec{s} \rangle + e_i \in \mathbb{Z}/q\mathbb{Z}$, and output $(\vec{a}_i, b_i)$.

The $\mathrm{LWE}_{n,q,\chi}$ assumption is that the $\mathrm{LWE}_{n,q,\chi}$ problem is hard.

For $n, q = \mathrm{poly}(\lambda)$, Regev gave a polynomial-time reduction from search LWE to decision LWE. Applebaum et al. [2] showed that the hardness of LWE is unaffected when the coefficients of secret $\vec{s}$ are chosen from the small error distribution χ.

Sometimes we prefer to view LWE in the following way. Let $\vec{c}_i = (b_i, \vec{a}_i)$ and $\vec{t} = (1, -\vec{s})$ for $b_i, \vec{a}_i, \vec{s}$ as above. Then $[\langle \vec{c}_i, \vec{t} \rangle]_q = [\vec{e}_i]_q$ is small for all i, where $[x]_q$ denotes the representative of x in $(-q/2, q/2]$. The LWE problem is to decide whether there exists a vector $\vec{t}$ that is "nearly orthogonal" to all of the $\vec{c}_i$'s.

Typically, χ is taken to be a discrete Gaussian distribution over $\mathbb{Z}$, with deviation $\sigma \ll q$. Rather than referring explicitly to the noise distribution χ, sometimes it is convenient to refer to a bound β on the size of the noise.

Definition 3.2 (β-bounded distributions). A distribution ensemble $\{\chi_n\}_{n \in \mathbb{N}}$, supported over the integers, is called β-bounded if $\Pr_{e \leftarrow \chi_n}[|e| > \beta] = \mathsf{negl}(n)$.

When the noise is extremely small or has some structure, there are sub-exponential algorithms to solve LWE [3]. For example, when $e_i \in \{0,1\}$ for all i, solving LWE is easy: taking tensor products, $\langle \vec{c}_i, \vec{t} \rangle \in \{0,1\}$ implies $\langle \vec{c}_i \otimes \vec{c}_i, \vec{t} \otimes \vec{t} \rangle - \langle \vec{c}_i, \vec{t} \rangle = 0$, giving us a $O(n^2)$-dimension error-free linear system to recover $\vec{t} \otimes \vec{t}$, hence $\vec{t}$. However, for discrete Gaussian error distributions with $\sigma = \text{poly}(n)$, the hardness of LWE stops depending so much on the noise bound β, and appears to depend more on the ratio q/β.

In particular, the LWE problem has been shown to be as hard *on average* (for random instances) as certain lattice problems *in the worst-case* (the hardest instances). A n-dimensional lattice is a (full-rank) additive subgroup of $\mathbb{R}^n$. For lattice dimension parameter n and number d, the shortest vector problem GapSVP_γ is the problem of distinguishing whether a n-dimensional lattice has a nonzero vector of Euclidean norm less than d or no nonzero vector shorter than $\gamma(n) \cdot d$. The gist of the theorem below is that if one can solve average-case n-dimensional LWE for ratio q/β then one can solve worst-case n-dimensional GAPSVP_γ for γ just a little larger than q/β.

Theorem 3.3 ([26, 27, 29, 31], Corollary 2.1 from [6]). *Let $q = q(n) \in \mathbb{N}$ be either a prime power or a product of small (size $\text{poly}(n)$) distinct primes, and let $\beta \geq \omega(\log n) \cdot \sqrt{n}$. Then there exists an efficient sampleable β-bounded distribution χ such that if there is an efficient algorithm that solves the average-case LWE problem for parameters n, q, χ, then:*

- *There is an efficient quantum algorithm that solves $\text{GapSVP}_{\tilde{O}(nq/\beta)}$ on any n-dimensional lattice.*
- *There is an efficient classical algorithm that solves $\text{GapSVP}_{\tilde{O}(nq/\beta)}$ on any n-dimensional lattice when $q \geq \tilde{O}(2^{n/2})$.*

Brakerski et al. [9] recently improved the classical result by removing the requirement on the size of q.

GAPSVP_γ is NP-hard for any constant γ, but unfortunately in cryptography we need γ to be larger (at least n in the theorem above). For $\gamma = \text{poly}(n)$, the fastest algorithm to solve GAPSVP_γ takes time $2^{O(n)}$. (As a crude rule of thumb, the fastest algorithm to solve GAPSVP_{2^k} takes roughly $2^{n/k}$ time [34].) Interestingly, there are no quantum algorithms for GAPSVP that perform significantly better than classical algorithms. In contrast, there are polynomial-time quantum algorithms for integer factorization and some other common problems used in cryptography.

4. Public key encryption from LWE

Regev [31] described a simple encryption scheme based on LWE. We describe a variant of his scheme here. We split key generation algorithm K into three parts Setup, $\mathsf{SecretKeyGen}$ and $\mathsf{PublicKeyGen}$. Let $[x]_q$ denote the integer $x \in (-q/2, q/2]$ that represents the coset of $x \in \mathbb{Z}/q\mathbb{Z}$.

- $\mathsf{Setup}(1^\lambda)$: Choose an odd integer modulus $q = q(\lambda)$, lattice dimension parameter $n = n(\lambda)$, and error distribution $\chi = \chi(\lambda)$ appropriately for LWE for security parameter λ. Also, choose parameter $m = m(\lambda) = O(n \log q)$. Let $params = (n, q, \chi, m)$.
- $\mathsf{SecretKeyGen}(params)$: Sample $\vec{s} \leftarrow \chi^n$. Set $\text{sk} = \vec{t} \leftarrow (1, -s_1, \ldots, -s_n) \in (\mathbb{Z}/q\mathbb{Z})^{n+1}$.

- PublicKeyGen$(params, \text{sk})$: Generate a matrix $A \leftarrow (\mathbb{Z}/q\mathbb{Z})^{m \times n}$ uniformly and a vector $\vec{e} \leftarrow \chi^m$. Set $\vec{b} = A \cdot \vec{s} + \vec{e}$. Set B to be the $(n+1)$-column matrix consisting of $\vec{b}$ followed by the n columns of A. Set the public key $\text{pk} = B$. (*Remark:* Observe that $B \cdot \vec{t} = \vec{e}$.)
- E$(params, \text{pk}, \mu)$: To encrypt message $\mu \in \{0,1\}$, sample uniform $\vec{r} \in \{0,1\}^m$, set $\vec{\mu} \leftarrow (\mu, 0, \ldots, 0) \in (\mathbb{Z}/q\mathbb{Z})^{n+1}$, and output the ciphertext:

$$\vec{c} \leftarrow \vec{\mu} + 2 \cdot \vec{r} \cdot B \in (\mathbb{Z}/q\mathbb{Z})^{n+1}.$$

- D$(params, \text{sk}, \vec{c})$: Output $[[\langle \vec{c}, \vec{t} \rangle]_q]_2$.

Decryption works correctly when the parameters are set so that $|\langle \vec{r}, \vec{e} \rangle| < q/4 - 1$ is guaranteed, since if $\vec{c} = \vec{\mu} + 2 \cdot \vec{r} \cdot B$ for $\mu \in \{0,1\}$, then $[\langle \vec{c}, \vec{t} \rangle]_q = [\mu + 2 \cdot \langle \vec{r}, \vec{e} \rangle]_q$ is an integer of magnitude $< q/2$ with the same parity as μ.

Interestingly, the encryption process of Regev's scheme already uses the fact the scheme is additively homomorphic. Each row $2 \cdot B_i$ of $2 \cdot B$ is an encryption of 0, in the sense that $[\langle B_i, \vec{t} \rangle]_q$ is small and even. To encrypt, one takes a random subset sum (defined by $\vec{r}$) of the $2 \cdot B_i$'s to obtain a "random" encryption of 0, and then one adds in $\vec{\mu}$ to get an encryption of μ.

This encryption process increases the size of the error: the error associated to the ciphertext is μ plus a subset sum of the errors associated to the $2 \cdot B_i$'s. One needs to set q large enough to "accommodate" the error expansion – again, one wants $|\mu + 2 \cdot \langle \vec{r}, \vec{e} \rangle| < q/2$ to ensure correct decryption.

The security of Regev's scheme follows from the following lemma [31].

Lemma 4.1 (Implicit in [31]). *Let $params = (n, q, \chi, m)$ be such that the* $\text{LWE}_{n,q,\chi}$ *assumption holds, with q odd. Then, for $m = O(n \log q)$ and B, $\vec{r}$ as generated above, the joint distribution $(B, 2 \cdot \vec{r} \cdot B)$ is computationally indistinguishable from uniform over $(\mathbb{Z}/q\mathbb{Z})^{m \times (n+1)} \times (\mathbb{Z}/q\mathbb{Z})^{n+1}$. Concretely, it suffices to take $m > 2n \log q$.*

The lemma says that, for Regev's encryption scheme, it is hard to distinguish a uniform matrix and uniform vector from a valid pk and a valid encryption of 0.

To sketch a proof of the lemma, observe that it follows from two claims: that it is hard to distinguish $(B, 2 \cdot \vec{r} \cdot B)$ from $(U, 2 \cdot \vec{r} \cdot U)$ where U is uniform in $(\mathbb{Z}/q\mathbb{Z})^{m \times (n+1)}$, and also $(U, 2 \cdot \vec{r} \cdot U)$ from $(U, \vec{u})$ where $\vec{u}$ is uniform in $(\mathbb{Z}/q\mathbb{Z})^{n+1}$. The first claim follows immediately from the LWE assumption, since given a LWE instance B or U, we can generate the $2 \cdot \vec{r} \cdot B$ or $2 \cdot \vec{r} \cdot U$ part ourselves. The second claim is true *statistically*. For large enough m, the distributions $(U, 2 \cdot \vec{r} \cdot U)$ and $(U, \vec{u})$ have negligible statistical distance from each other when q is odd.

Now, let us use the lemma to reduce LWE to the semantic security of Regev's encryption scheme. Assume an adversary wins the semantic security game with non-negligible advantage. We imagine two games between the challenger and the adversary. In Game 0, the challenger uses the distribution $(B, 2 \cdot \vec{r} \cdot B)$ to generate its public key $\text{pk} = B$ and challenge ciphertext $\vec{c} \leftarrow \vec{\mu} + 2 \cdot \vec{r} \cdot B$. By assumption, the adversary guesses μ with non-negligible advantage. In Game 1, uses uniform $(U, \vec{u}) \in (\mathbb{Z}/q\mathbb{Z})^{m \times (n+1)} \times (\mathbb{Z}/q\mathbb{Z})^{n+1}$, sets $\text{pk} = U$, and sets $\vec{c} \leftarrow \vec{\mu} + \vec{u}$. In Game 1, since $\vec{u}$ is uniform, the adversary has no advantage guessing μ. We guess that the distribution is $(B, 2 \cdot \vec{r} \cdot B)$ (that we are in Game 0) if the adversary guesses μ correctly; otherwise, we guess the distribution is uniform (that we are in Game 1).

One can show that if the adversary guesses μ correctly in Game 0 with probability $1/2 + \epsilon$, then we guess the distribution correctly with probability $1/2 + \epsilon/2$.

5. Leveled FHE from LWE

The Gentry-Sahai-Waters (GSW) leveled FHE scheme [20] is currently the conceptually simplest FHE scheme whose security is based on LWE. As a warm-up to build intuition, we first describe how a noise-free (but insecure) version of GSW would work. Then, we introduce noise, describe how to fix the problems it causes, and reduce the security of GSW to the security of Regev's scheme (hence to LWE).

5.1. Thought experiment: Leveled FHE from learning *without* errors. Imagine that Regev's encryption scheme had no error, that an encryption of $\mu \in \{0, 1\}$ is simply a vector $\vec{c} \in (\mathbb{Z}/q\mathbb{Z})^{n+1}$ such that $\langle \vec{c}, \vec{t} \rangle = \mu \in \mathbb{Z}/q\mathbb{Z}$, where $\vec{t}$ is the secret key. How can we add and multiply such ciphertexts so as to add and multiply the plaintexts inside?

Addition is easy. Given two ciphertexts $\vec{c}_1, \vec{c}_2$ that happen to encrypt μ_1, μ_2, we add them to obtain a ciphertext that encrypts the sum: $\langle \vec{c}_1 + \vec{c}_2, \vec{t} \rangle = \mu_1 + \mu_2$.

Multiplication is tricker. We can use tensor products: $\langle \vec{c}_1 \otimes \vec{c}_2, \vec{t} \otimes \vec{t} \rangle = \mu_1 \cdot \mu_2$. However, then each circuit level of multiplications squares the dimension of the ciphertexts, making the scheme non-compact and inefficient.

To get compact multiplication, a better idea is to use matrix multiplication. Specifically, let an encryption of μ be a square matrix C such that $C \cdot \vec{t} = \mu \cdot \vec{t}$. In other words, the secret key is an *eigenvector* of the ciphertext matrix, and the message is the eigenvalue.[3] Addition and multiplication of ciphertexts induces addition and multiplication of plaintexts (eigenvalues). Decryption is a ring homomorphism from the ring of matrices having $\vec{t}$ as an eigenvector to the corresponding eigenvalue.

Unfortunately, this scheme is easy to attack. The encryptions of 0 form a subspace that is easily identified (via linear algebra) once enough encryptions of 0 are collected. More broadly, this eigenvector-based FHE scheme falls within the so-far-unsuccessful *hidden ring homomorphism* approach to FHE. In this approach, the message space $\mathcal{M}$ and ciphertext space $\mathcal{C}$ are rings, and decryption $\mathsf{D}_{\mathrm{sk}} : \mathcal{C} \to \mathcal{M}$ is a ring homomorphism that depends on the secret key sk. Addition and multiplication of ciphertexts induce addition and multiplication of plaintexts. Encryptions of 0 form an *ideal* $\mathcal{I}$ in $\mathcal{C}$, while encryptions of 1 are in $1 + \mathcal{I}$. Semantic security relies on the hardness of the *ideal membership problem*: roughly, distinguish whether an element of $\mathcal{C}$ is in $\mathcal{I}$. Another example in this framework is the Polly Cracker scheme proposed by Fellows and Koblitz [13], where the secret key is a secret point in $\vec{s} \in \mathbb{F}_q^n$, and μ is encrypted as a "random" multivariate polynomial that evaluates to μ at $\vec{s}$. Unfortunately, so far, there are no FHE schemes based on hidden ring homomorphisms that are both compact and secure (though the approach has not been ruled out).

5.2. Error-Preserving transformations. As we will see, the GSW scheme uses exactly the above eigenvector approach, but adds noise to it. In GSW, the secret key is a vector $\vec{v}$ with a special form, and an encryption of μ is a matrix C such that $C \cdot \vec{v} = \mu \cdot \vec{v} + \vec{e}$ for small error vector $\vec{e}$ – that is, $\vec{v}$ is an *approximate eigenvector* of the ciphertext, with the message

[3] Note that since we work modulo q, eigenvectors here do not have the usual geometric interpretation.

as the eigenvalue. The noise makes multiplication tricky again, since

$$C_1 \cdot C_2 \cdot \vec{v} = C_1 \cdot (\mu_2 \cdot \vec{v} + \vec{e}_2) = \mu_1 \cdot \mu_2 \cdot \vec{v} + (\mu_2 \cdot \vec{e}_1 + C_1 \cdot \vec{e}_2).$$

The new noise $\mu_2 \cdot \vec{e}_1 + C_1 \cdot \vec{e}_2$ depends not only on the old noises, but also on the second message and the first ciphertext. To ensure that the magnitude of the noise grows at most by a polynomial factor with each circuit level of multiplication, we need to keep the messages small (we do this by restricting messages to $\{0,1\}$ and using NAND gates) and also keep the ciphertexts small.

Here, we describe embarrassingly simple (but very useful) error-preserving transformations that an Evaluator can apply to make the entries of a ciphertext matrix small (in $\{0,1\}$) without knowing or altering what the ciphertext encrypts. The idea is simply to use binary decomposition: we decompose each mod-q coefficient into $\log_2 q$ coefficients in $\{0,1\}$.

Specifically, let $\vec{c}$, $\vec{t}$ be vectors in $(\mathbb{Z}/q\mathbb{Z})^k$. Let $\ell = \lfloor \log_2 q \rfloor + 1$ and $N = k \cdot \ell$. Let $\mathsf{BitDecomp}(\vec{c}) = (c_{1,0}, \ldots, c_{1,\ell-1}, \ldots, c_{k,0}, \ldots, c_{k,\ell-1})$, a N-dimensional vector where $c_{i,j}$ is the j-th bit in c_i's binary representation, bits ordered least significant to most significant. For $\vec{c}^* = (c_{1,0}, \ldots, c_{1,\ell-1}, \ldots, c_{k,0}, \ldots, c_{k,\ell-1})$, let $\mathsf{BitDecomp}^{-1}(\vec{c}^*) = (\sum 2^j \cdot c_{1,j}, \ldots, \sum 2^j \cdot c_{k,j})$ be the inverse of BitDecomp, but well-defined even when the input is not a 0/1 vector. For N-dimensional $\vec{c}^*$, let $\mathsf{Flatten}(\vec{c}^*) = \mathsf{BitDecomp}(\mathsf{BitDecomp}^{-1}(\vec{c}^*))$, a N-dimensional vector with $0/1$ coefficients. When A is a matrix, let $\mathsf{BitDecomp}(A)$, $\mathsf{BitDecomp}^{-1}(A)$, or $\mathsf{Flatten}(A)$ be the matrix formed by applying the operation to each row of A separately. Finally, let $\mathsf{Powersof2}(\vec{t}) = (t_1, 2t_1, \ldots, 2^{\ell-1}t_1, \ldots, t_k, 2t_k, \ldots, 2^{\ell-1}t_k)$, a N-dimensional vector. Here are some obvious facts:

- $\langle \vec{c}, \vec{t} \rangle = \langle \mathsf{BitDecomp}(\vec{c}), \mathsf{Powersof2}(\vec{t}) \rangle$.
- For any N-dimensional $\vec{c}^*$:
  $\langle \vec{c}^*, \mathsf{Powersof2}(\vec{t}) \rangle = \langle \mathsf{BitDecomp}^{-1}(\vec{c}^*), \vec{t} \rangle = \langle \mathsf{Flatten}(\vec{c}^*), \mathsf{Powersof2}(\vec{t}) \rangle$.

In the GSW scheme, which we finally formally describe in the next subsection, we will use $\vec{v} \leftarrow \mathsf{Powersof2}(\vec{t})$ as the secret key vector, rather than $\vec{t}$. The salient feature of Flatten is that we can apply it to a matrix C that encrypts a message under $\mathsf{Powersof2}(\vec{t})$ without affecting its product with $\mathsf{Powersof2}(\vec{t})$ and hence what it encrypts, and (importantly) without knowing $\vec{t}$. By Flattening ciphertexts after each operation, we ensure that the next operation will increase the magnitude of the error by only a polynomial factor.

5.3. The GSW leveled FHE scheme from LWE. Brakerski and Vaikuntanathan were the first to construct a leveled FHE scheme based on LWE [10]. However, the scheme by Gentry, Sahai and Waters is particularly simple. It uses a "compiler" that transforms any LWE-based public-key encryption scheme $(\mathsf{K}, \mathsf{E}, \mathsf{D})$ that has certain natural properties into a LWE-based leveled FHE scheme $(\mathsf{GSW.K}, \mathsf{GSW.E}, \mathsf{GSW.D}, \mathsf{GSW.NAND})$ capable of Evaluating circuits of NAND gates. Regev's scheme has the needed properties. The properties are:

1. **Property 1 (Vectors and parameters)**: The ciphertext and decryption key are vectors $\vec{c}, \vec{t} \in (\mathbb{Z}/q\mathbb{Z})^{n'}$ for some n'. The first coefficient of $\vec{t}$ is 1 and q is odd.
2. **Property 2 (Small dot product)**: If $\vec{c}$ encrypts 0, then $\langle \vec{c}, \vec{t} \rangle$ is "small".
3. **Property 3 (Security)**: Encryptions of 0 are indistinguishable from uniform vectors over $\mathbb{Z}/q\mathbb{Z}$ (under LWE).

The parameters n, q, χ of the underlying encryption scheme determine the depth L of the circuits that GSW can Evaluate. So, the compiler is not completely black box; K must be tweaked to depend on L. (We will discuss how L affects parameter sizes later.) The GSW scheme works as follows.

- $\mathsf{GSW.K}(1^\lambda, 1^L)$: Compute $\mathsf{K}(1^\lambda, 1^L)$ to obtain parameters $params$, secret vector $\mathrm{sk} = \vec{t} \in (\mathbb{Z}/q\mathbb{Z})^{n'}$ and public key pk. Let $\ell = \lfloor \log q \rfloor + 1$ and $N = n' \cdot \ell$. Set $\vec{v} = \mathsf{Powersof2}(\vec{t})$.
- $\mathsf{GSW.E}(params, \mathrm{pk}, \mu \in \{0,1\})$: Set $\vec{c}'_i \leftarrow \mathsf{E}(params, \mathrm{pk}, 0)$ for i from 1 to N. (*Remark:* These are just N encryptions of 0 under the public key encryption scheme.) Set $C' \in (\mathbb{Z}/q\mathbb{Z})^{N \times n'}$ to be the matrix with rows $\{\vec{c}'_i\}$. Output the ciphertext C given below. (I_N is the N-dimensional identity matrix.)

$$C = \mathsf{Flatten}\big(\mu \cdot I_N + 2 \cdot \mathsf{BitDecomp}(C')\big) \in (\mathbb{Z}/q\mathbb{Z})^{N \times N}.$$

- $\mathsf{GSW.D}(params, \mathrm{sk}, C)$: Let $\vec{c}_1$ be the first row of C. Output $[[\langle \vec{c}_1, \vec{v} \rangle]_q]_2$.
- $\mathsf{NAND}(C_1, C_2)$: To NAND two ciphertexts $C_1, C_2 \in (\mathbb{Z}/q\mathbb{Z})^{N \times N}$, output $\mathsf{Flatten}(I_N - C_1 \cdot C_2)$.

Decryption works, since if C is as above, then

$$\begin{aligned} C \cdot \vec{v} &= (\mu \cdot I_N + 2 \cdot \mathsf{BitDecomp}(C')) \cdot \vec{v} \ \text{[Flatten preserves product with } \vec{v}\text{]} \\ &= \mu \cdot \vec{v} + 2 \cdot C' \cdot \vec{t} \qquad \text{[}\mathsf{BitDecomp}(C') \cdot \mathsf{Powersof2}(\vec{t}) = C' \cdot \vec{t}\text{]} \\ &= \mu \cdot \vec{v} + 2 \cdot small \qquad \text{[By Property 2 above].} \end{aligned}$$

Since $v_1 = 1$, the integer $[\langle \vec{c}_1, \vec{v} \rangle]_q = \mu \cdot v_1 + 2 \cdot small$ is small and has the same parity as μ, allowing recovery of $\mu \in \{0,1\}$ when $|small| < q/4 - 1$.

NAND works, since if C_1, C_2 happen to be valid encryptions of $\mu_1, \mu_2 \in \{0,1\}$ with errors $\vec{e}_1, \vec{e}_2$, then:

$$\mathsf{NAND}(C_1, C_2) \cdot \vec{v} = (I_N - C_1 \cdot C_2) \cdot \vec{v} = (1 - \mu_1 \cdot \mu_2) \cdot \vec{v} - \mu_2 \cdot \vec{e}_1 - C_1 \cdot \vec{e}_2$$

Note that NAND maintains the invariant that if the input messages are in $\{0,1\}$, then so is the output message. With this invariant, and using Flatten to ensure that C_1's coefficients are in $\{0,1\}$, the output error is at most $N + 1$ times larger than the bigger input error.

Theorem 5.1. GSW *is semantically secure under the LWE assumption.*

Proof. By Property 3, C' is indistinguishable from a uniform matrix under LWE. Thus, since q is odd, $\mathsf{BitDecomp}^{-1}(C) = \mu \cdot \mathsf{BitDecomp}^{-1}(I_N) + 2 \cdot C'$ is indistinguishable from uniform U. But then $C = \mathsf{Flatten}(C)$ is indistinguishable from $\mathsf{BitDecomp}(U)$, where the latter is independent of μ. □

5.4. Parameters and performance. Suppose that we would like to use GSW to evaluate NAND circuits with up to L levels. How should we set the parameters to ensure correctness and security? How much computation does the Evaluate algorithm use per NAND gate?

We have seen that each NAND gate multiplies the magnitude of the error by a factor of at most $N + 1$. If β is a bound on the error magnitude of *fresh* ciphertexts, then L levels

of NAND gates amplify the error magnitude to at most $\beta \cdot (N+1)^L$. Decryption works correctly despite such large error, as long as $q/4 - 1 > \beta \cdot (N+1)^L \rightsquigarrow q/\beta > 4N^L$. The ratio q/β must grow exponentially with L to "accommodate" the noise.

Using the rule of thumb that solving GAPSVP$_{2^k}$ in n-dimensional lattices takes time roughly $2^{n/k}$, and acknowledging that a GAPSVP$_{q/\beta}$ solver would break the scheme, the lattice dimension n (hence N) must increase linearly with $\log(q/\beta)$ to maintain fixed 2^λ security against known attacks. But let us brush this issue under the rug and view n as a fixed parameter. Choosing χ so that β is not too large, and since in practice there is no reason to have $\log q$ grow super-linearly with n, we have $\log q = O(L \log N) = O(L(\log n + \log\log q)) = O(L \log n)$. Given that the NAND procedure is dominated by multiplication of two $N \times N$ matrices for $N = O(n \log q) = \tilde{O}(nL)$, we have the following theorem to characterize the performance of GSW.

Theorem 5.2. *For dimension parameter n and depth parameter L,* GSW *correctly evaluates depth-L circuits of* NAND *gates with $\tilde{O}((nL)^\omega)$ field operations per gate, where $\omega < 2.3727$ is the matrix multiplication exponent.*

Thus, we obtain a leveled FHE scheme with poly(λ, L) computation per NAND gate that achieves 2^λ security against known attacks.

However, even the most theoretical mathematician or computer scientist should be able to see that this scheme will be too slow in practice to Evaluate even moderately complex functions. While LWE-based GSW is far from being the fastest FHE scheme, a big open problem remains: construct a FHE scheme that is truly practical!

As described so far, GSW may leave even a theoretician unsatisfied, as it leaves ample room for qualitative improvement. It begs some questions: Can we make per-gate computation independent of L? Can we Evaluate a priori unbounded depth circuits? Can we actually reduce the noise rather than merely "accommodating" it? For example, can we devise a "refresh" procedure that reduces the noise level of a ciphertext without altering what it encrypts, so that we can Evaluate ad infinitum, refreshing when needed?

The theoretician, at least, may find some solace in the answers we provide in the next section, where we describe precisely such a "refresh" procedure, called *bootstrapping*, that allows Evaluation of unbounded-depth circuits with per-gate computation independent of the depth.

6. Bootstrapping: Homomorphic encryption for unbounded depth circuits

In GSW and all current FHE schemes, ciphertexts are "noisy". Computing over the ciphertexts increases the noise, until eventually the noise becomes bigger than the modulus q, and all hope of reliably decrypting the message correctly is lost. Must we surrender to this life-destroying entropy? Or is there some way to "rejuvenate" an old noisy ciphertext, to create a new ciphertext that encrypts the same value but with much less noise, so that it can safely participate in more computation? Here, we describe a procedure called *bootstrapping* that refreshes ciphertexts, gives them a sort of immortality, so that we can Evaluate unbounded depth circuits with per-gate computation independent of the depth.

6.1. Self-Referentiality in encrypted computation. Can the brain understand itself? Philosophically, it seems appealing to think that, as a brain becomes more complex, so does the

task of understanding it, so that self-understanding remains eternally just out of reach.

Here we consider a somewhat similar question: Can a homomorphic encryption scheme decrypt itself? The decryption function of a homomorphic encryption scheme is, after all, just another function that we can try to plug into the Evaluate algorithm. But does it work? Or, is it the case that, for any L, the decryption function of a leveled FHE scheme capable of Evaluating depth-L circuits has depth greater than L, beyond the Evaluation capacity of scheme?

This is no idle brain-teaser. Actually, among the functions than an FHE scheme can Evaluate, its own decryption function is not only the most interesting, but perhaps also the most useful. Let us consider what we can do with such self-referential encrypted computation. Suppose c encrypts μ under $(\mathrm{pk}, \mathrm{sk})$. Set $\overline{\mathrm{sk}_i} \leftarrow \mathsf{E}(\mathrm{pk}, \mathrm{sk}_i)$ for all of the bits $\{\mathrm{sk}_i\}$ of sk – that is, the ciphertexts $\{\overline{\mathrm{sk}_i}\}$ are an *encryption of the key under itself*. We will publish this encryption of the secret key, so that Evaluators can use it. Set $\overline{c_i} \leftarrow \mathsf{E}(\mathrm{pk}, c_i)$ for all of the bits $\{c_i\}$ of c – that is, these ciphertexts are a *double encryption* of μ. Now, suppose that the leveled FHE scheme can correctly Evaluate L levels, but Evaluating the decryption function D requires only $L-1$ levels. Consider the following ciphertext:

$$c' \leftarrow \mathsf{V}(\mathrm{pk}, \mathsf{D}, (\{\overline{\mathrm{sk}_i}\}, \{\overline{c_i}\})).$$

By the correctness of Evaluate:

$$\mathsf{D}(\mathrm{sk}, c') = \mathsf{D}(\{\mathrm{sk}_i\}, \{c_i\}) = \mu.$$

That is, the new ciphertext c' encrypts the same value as the old ciphertext c. (Interestingly, Evaluating the decryption function on the double encryption $\{\overline{c_i}\}$ removes the *inner* encryption.) Moreover, since c' is the result of Evaluating a circuit of only $L-1 < L$ levels on *fresh* ciphertexts $\{\overline{\mathrm{sk}_i}\}, \{\overline{c_i}\}$, it (possibly unlike c) can be used safely as input to one more NAND gate. Of course, an Evaluator can use this refreshing trick as often as necessary to ensure the noise level of the ciphertexts remains safely bounded. In short, if we have a magical homomorphic encryption scheme capable of Evaluating its own decryption circuit with room to spare, then that homomorphic encryption scheme can be *bootstrapped* into a pure FHE scheme capable of evaluating unbounded depth circuits.[4]

More formally, Gentry [16] defined and proved the following.

Definition 6.1 (Bootstrappable encryption scheme). A homomorphic encryption scheme $\mathcal{E}$ is called *bootstrappable* if $\mathcal{E}$ compactly evaluates all circuits of depth at most $(D+1)$, where D is the depth of $\mathcal{E}$'s decryption circuit, and the computational complexity of $\mathcal{E}$'s algorithms is polynomial in the security parameter and (in the case of the evaluation algorithm) the size of the circuit.

Theorem 6.2 (Bootstrapping Theorem). *For any bootstrappable encryption scheme $\mathcal{E}$, there exists a leveled FHE scheme $\{\mathcal{E}^{(L)}\}$ with related security.*

Letting S be the size of $\mathcal{E}$'s decryption circuit, the per-gate evaluation complexity of the leveled FHE is exactly the complexity of evaluating a $(2S+1)$-gate circuit using the bootstrappable scheme: independent of the depth of the circuit.

Under an assumption of circular security *– that is, an assumption that semantic security is preserved despite publishing an encryption of the secret key under its corresponding public key – one obtains a pure FHE scheme.*

[4] For more intuition, see Gentry's (somewhat dated) 2010 survey [17] on FHE for a full-fledged physical analogy for bootstrapping in terms of gloveboxes inside gloveboxes.

Gentry also provided the first bootstrappable and fully homomorphic encryption schemes based on plausible assumptions.

Circular encryptions sound dangerous, but for most encryption schemes it appears that revealing an encryption of sk under pk does not lead to any attack. On the other hand, it is typically difficult to *prove* that an encryption scheme is circular-secure, hence the need for the additional assumption.

To avoid the circular-security assumption, one can instead provide an *acyclic chain* of encrypted secret keys. One generates a key pair $(\mathrm{pk}_i, \mathrm{sk}_i)$ for each level of the circuit, and provides an encryption of sk_i under pk_{i+1}. In this case, one can prove that the encrypted secret key bits are indistinguishable from encryptions of 0 as long as $\mathcal{E}$ is semantically secure.

6.2. Evaluating the GSW decryption circuit. So, can GSW decrypt itself? It turns out it can, but we need one more trick. The concept of the trick is that, before we bootstrap, we can pre-process the ciphertext into a form that does not permit any more homomorphic operations, but is much less complex to decrypt (and hence to bootstrap).

In more detail, recall that a GSW ciphertext is a matrix, but we use only the first row of the matrix during decryption: $\mu = [[\langle \vec{c}_1, \vec{v} \rangle]_q]_2$. Also, we can use $\vec{t}$ rather than $\vec{v} = \mathsf{Powersof2}(\vec{t})$ as the secret key: $\mu = [[\langle \mathsf{BitDecomp}^{-1}(\vec{c}_1), \vec{t} \rangle]_q]_2$. Now, we only need to decrypt (bootstrap) $\mathsf{BitDecomp}^{-1}(\vec{c}_1)$. However, there is still a problem: the complexity of decrypting it depends on q and hence on L, the number of levels the scheme can Evaluate. Can we remove this dependence, to obtain a ciphertext whose decryption complexity is polynomial in the security parameter λ and completely independent of L? If so, then we are done.

Brakerski and Vaikuntanathan [10] gave a particularly clean way of removing this dependence. They showed that we can apply *modulus reduction* and *dimension reduction* to a Regev-type ciphertext $\vec{c}$ (like our $\mathsf{BitDecomp}^{-1}(\vec{c}_1)$ above), so that the complexity of decrypting the final ciphertext becomes independent of L. Modulus reduction takes an initial ciphertext $\vec{c}$ that encrypts μ modulo q, and outputs a new ciphertext that encrypts μ modulo a smaller modulus p. Dimension reduction reduces the dimension of the ciphertext vector. After applying modulus and dimension reduction, we obtain a ciphertext $\vec{c}^*$ of $\mathrm{poly}(\lambda)$ dimension such that $\mu = [[\langle \vec{c}^*, \vec{t} \rangle]_p]_2$ for small p (e.g., p may even be only polynomial in the security parameter). The size of $\vec{c}^*$ is independent of L.

Let us sketch how modulus reduction works. (We omit a description of dimension reduction.) Recall that Applebaum et al. [2] showed that the hardness of LWE is unaffected when the coefficients of secret key are chosen from the small error distribution χ. When $\vec{t}$ is small and $[\langle \vec{c}_i, \vec{t} \rangle]_q$ is small, then $[\langle \vec{c}_i^*, \vec{t} \rangle]_p$ is also small, where $\vec{c}_i^* = \lfloor (p/q) \cdot \vec{c}_i \rceil$ is simply p/q times $\vec{c}_i$ rounded. The following easy lemma makes this more precise, and also shows that we can preserve other aspects of the noise, such as its parity.

Lemma 6.3. *Let p and q be two odd moduli, and let $\vec{c}$ be an integer vector. Define $\vec{c}^*$ to be the integer vector closest to $(p/q) \cdot \vec{c}$ such that $\vec{c}^* = \vec{c} \bmod 2$. Then, for any $\vec{t}$ with $|[\langle \vec{c}, \vec{t} \rangle]_q| < q/2 - (q/p) \cdot \ell_1(\vec{t})$, we have*

$$\begin{aligned} [\langle \vec{c}^*, \vec{t} \rangle]_p &= [\langle \vec{c}, \vec{t} \rangle]_q \bmod 2 \quad \text{and} \\ |[\langle \vec{c}^*, \vec{t} \rangle]_p| &< (p/q) \cdot |[\langle \vec{c}, \vec{t} \rangle]_q| + \ell_1(\vec{t}) \end{aligned}$$

where $\ell_1(\vec{t}) = \sum |t_i|$ is the ℓ_1-norm of $\vec{t}$.

Proof. For some integer k, we have $[\langle \vec{c}, \vec{t} \rangle]_q = \langle \vec{c}, \vec{t} \rangle - kq$. For the same k, let $e_p = \langle \vec{c}^*, \vec{t} \rangle - kp \in \mathbb{Z}$. Since $\vec{c}^* = \vec{c}$ and $p = q$ modulo 2, we have $e_p = [\langle \vec{c}, \vec{t} \rangle]_q \bmod 2$. To finish the proof, it suffices to prove that $e_p = [\langle \vec{c}^*, \vec{t} \rangle]_p$ and that it has small enough norm. We have $e_p = (p/q)[\langle \vec{c}, \vec{t} \rangle]_q + \langle \vec{c}^* - (p/q)\vec{c}, \vec{t} \rangle$, and therefore $|e_p| \leq (p/q)[\langle \vec{c}, \vec{t} \rangle]_q + \ell_1(\vec{t}) < p/2$. The latter inequality implies $e_p = [\langle \vec{c}^*, \vec{t} \rangle]_p$. □

An alternative view of modulus reduction is that we might as well divide $\vec{c}$ by q and consider its dot product with $\vec{t}$ *modulo 1* – the q merely represents the fact that we represent coefficients of $\vec{c}$ with $\log q$ bits of precision. When we begin to Evaluate a deep circuit, we need lots of precision, since many noise-increasing operations remain. But as we complete the circuit, we can drop precision, allowing the ciphertext to become smaller.

To make a homomorphic encryption scheme bootstrappable, one merely sets the parameters of the scheme so that it is capable of Evaluating the reduced decryption circuit (plus one more NAND gate). The reduced decryption circuit has depth logarithmic in the security parameter. Since each level of NAND gates increases the noise by a polynomial factor, we can bootstrap GSW by setting q to be quasi-polynomial, and (modulo the circular security issue) we can base the security of GSW on LWE for quasi-polynomial factors. Very recently, Brakerski and Vaikuntanathan [11] showed how to go from quasi-polynomial to polynomial by devising a decryption algorithm that, when Evaluated with GSW, increases the noise by only a polynomial factor.

7. Looking beyond bootstrapping

In some sense, the current approach to FHE using noise and bootstrapping has been enormously successful. As we have seen, we can Evaluate arbitrary encrypted functions over encrypted data with overhead only polynomial in the security parameter, independent of the complexity of the function. In fact, we can do even better. We can pack many plaintexts into each ciphertext, and perform batch SIMD (simultaneous instruction multiple data) on encrypted arrays, so as to Evaluate a function many times in parallel without additional computation over many encrypted data-sets [7, 8, 19, 35]. Using a variant of LWE called ring LWE in which the coefficient vectors are over the ring of integers of a cyclotomic number field, we can even move data in encrypted arrays between different array "slots" by using automorphisms of the ring. Using ring LWE with ciphertext-packing and automorphisms, we can get the overhead of FHE down to *polylogarithmic* in the security parameter [19].

Unfortunately, it turns out that polylogarithmic can still be impractically large. The overhead of current FHE schemes is still at least in the high millions for reasonable values of the security parameter. The problem is noise and bootstrapping: Evaluating the decryption circuit after Evaluating each NAND gate in our function seems to inherently require huge overhead, even it is batched to refresh multiple ciphertexts simultaneously. Can we do better? Can we eliminate bootstrapping, or even eliminate noise altogether?

7.1. Can we refresh ciphertexts without bootstrapping? Bootstrapping reduces the noise of a ciphertext by applying Decryption to it inside an Evaluation. But is there a more *direct* way to reduce the noise so as to Evaluate unbounded depth circuits? This is a fascinating open problem.

Quantum error correction (QEC) has a high-level similarity to ciphertext refreshing. To

correct noise in a quantum computation (e.g., phase errors in the qubits), QEC introduces some ancillary bits to the computation, uses them to compute an error correction syndrome over the primary qubits, measures the syndrome, and uses the result to adjust the quantum state of the primary qubits. A peculiarity of QEC is that measurement of the ancillary bits must not reveal anything about the correct values of the primary bits; else, the measurement would collapse the computation. Can we construct an analogous noise reduction technique for FHE, where an Evaluator can compute a syndrome that allows it to reduce ciphertext noise, but still cannot learn what the ciphertext encrypts?

Tao's computational program for Navier-Stokes [36] might be another place to look for new ideas to reduce ciphertext noise. Part of the reason bootstrapping is slow is that it goes "outside of the system": it refreshes a ciphertext not by acting on it directly, but rather by using the ciphertext to construct a function that is Evaluated over fresh encryptions of the secret key bits. If, instead, we could manage ciphertext noise *endogenously* (like Tao's water-based circuits), one could hope that eliminating the layer of indirection would also reduce computational complexity.

7.2. Can we eliminate noise altogether? The noisiness of LWE-based ciphertexts is the basis of their security, but also an obstacle to making FHE practical. Can we construct an FHE scheme without noise?

Without noise, decryption in GSW is a purely linear function, and the system can be broken easily using linear algebra. More generally, for any encryption scheme in which $\mathsf{D}(\mathrm{sk}, c)$ is a degree k polynomial, we can view $\mathsf{D}(\mathrm{sk}, c)$ as a dot product $\langle M(\mathrm{sk}), M(c)\rangle$ of the vectors of monomials of degree at most k associated to sk and c, and an attacker can use linear algebra to break semantic security in time $\lambda^{O(k)}$. So, to get 2^λ security, the degree of $\mathsf{D}(\mathrm{sk}, c)$ must be essentially linear in λ, a "complex" function. And yet, for an FHE scheme, $\mathsf{D}(\mathrm{sk}, c)$ must also be robust and flexible enough to allow computation.

Interestingly, the noise in LWE-based schemes boosts the degree of the decryption function. Although $[[\vec{c}, \vec{t}]_q]_2$ looks "almost linear", the rounding makes it high degree both modulo q and modulo 2. On the other hand, the "almost linearity" of decryption allows computation.

As some some final food for thought, we sketch an interesting but so-far-unrealized framework due to Nuida [28] for constructing pure noise-free FHE using non-abelian groups. Unfortunately, the framework also illustrates the difficulty of avoiding linear algebra attacks, even in contexts (using groups rather than rings) where one might hope they are inapplicable. First, a couple of definitions:

Definition 7.1 (Perfect Group Pairs). We call (G, H) a *perfect group pair* if G and H are both finite perfect groups (equal to their commutator subgroups) and H is a normal subgroup of G. We also require that G and H have efficient $(\mathrm{polylog}(|G|))$ operations – in particular, given a set of group generators of G or H, one can re-randomize them to obtain a random set of B group elements (for some polynomial bound B) that generate the same group.

Definition 7.2 (Perfect Group Pair Decision (PGP) Problem). Given (generators for) a perfect group pair G and H, and a third set of generators that generates G or H, distinguish which.

The form of the ciphertexts is simple: an encryption of 1 is a set of generators of G, while an encryption of 0 is a set of generators of H. The public key contains encryptions of '1' and '0' that the encrypter can randomize to generate its ciphertext. Decryption will

use some (unspecified) secret key $\tau_{G,H}$ that allows the keyholder to distinguish between generators for G and H. Semantic security follows directly from the PGP assumption and the re-randomizability of the group generators.

We describe homomorphic operations only for AND and OR gates (monotone circuits). Suppose the inputs to the gate are generators of (unknown) groups K_1, K_2. To Evaluate an OR gate, output (randomized) generators for the join of K_1 and K_2. (The output group is G iff an input group is G and H otherwise, and thus computes OR correctly.) To Evaluate an AND gate, output (randomized) generators for the commutator $[K_1, K_2]$. (Since H is normal in G, the output group is H iff an input group is H and G otherwise, and thus computes AND correctly.)

The main open problem for this framework is to find suitable perfect group pairs. It is easy to find perfect group pairs for which the PGP problem is easy: for example, take $G = H \times K$ for perfect groups H and K, where the extra coordinate makes elements of G easy to identify. Also, there are various perfect matrix group pairs (G, H) where the PGP problem is less trivial, but still ultimately solvable via linear algebra. Even if the groups are not initially presented as matrices, one must avoid groups with efficiently computable representations that enable linear algebra attacks. Still, this framework, though unrealized, serves as a useful counterpoint to the notion that noise and bootstrapping may be necessary to obtain FHE.

References

[1] Daniel Apon, Xiong Fan, Jonathan Katz, Feng-Hao Liu, Elaine Shi, and Hong-Sheng Zhou, *Non-interactive cryptography in the ram model of computation*, IACR Cryptology ePrint Archive, 2014:154, 2014.

[2] Benny Applebaum, David Cash, Chris Peikert, and Amit Sahai, *Fast cryptographic primitives and circular-secure encryption based on hard learning problems*, In CRYPTO, Springer, 2009, pp. 595–618.

[3] Sanjeev Arora and Rong Ge, *New algorithms for learning in presence of errors*, In ICALP, Springer, 2011, pp. 403–415.

[4] Boaz Barak, Oded Goldreich, Russell Impagliazzo, Steven Rudich, Amit Sahai, Salil P. Vadhan, and Ke Yang, *On the (im)possibility of obfuscating programs*, In CRYPTO, Springer, 2001, pp. 1–18.

[5] Dan Boneh, Amit Sahai, and Brent Waters, *Functional encryption: Definitions and challenges* In TCC, Springer, 2011, pp. 253–273.

[6] Zvika Brakerski, *Fully homomorphic encryption without modulus switching from classical gapsvp*, In CRYPTO, Springer, 2012, pp. 868–886.

[7] Zvika Brakerski, Craig Gentry, and Shai Halevi, *Packed ciphertexts in lwe-based homomorphic encryption* In PKC, Springer, 2013, pp. 1–13.

[8] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan, *(leveled) fully homomorphic encryption without bootstrapping*, In ITCS, ACM, 2012, pp, 309–325.

[9] Zvika Brakerski, Adeline Langlois, Chris Peikert, Oded Regev, and Damien Stehlé, *Classical hardness of learning with errors*, In STOC, ACM, 2013, pp. 575–584.

[10] Zvika Brakerski and Vinod Vaikuntanathan, *Efficient fully homomorphic encryption from (standard) LWE*, In FOCS, IEEE, 2011, pp. 97–106.

[11] Zvika Brakerski and Vinod Vaikuntanathan, *Lattice-based fhe as secure as pke*, In ITCS, ACM, 2014, pp. 1–12.

[12] Whitfield Diffie and Martin E. Hellman, *New directions in cryptography*, IEEE Transactions on Information Theory **22**(6) (1976), 644–654.

[13] Michael Fellows and Neal Koblitz, *Combinatorial cryptosystems galore!*, Contemporary Mathematics **168** (1994), 51–51.

[14] Sanjam Garg, Craig Gentry, and Shai Halevi, *Candidate multilinear maps from ideal lattices*, In EUROCRYPT, Springer, 2013, pp. 1–17.

[15] Sanjam Garg, Craig Gentry, Shai Halevi, Mariana Raykova, Amit Sahai, and Brent Waters, *Candidate indistinguishability obfuscation and functional encryption for all circuits*, In FOCS, IEEE, 2013, pp. 40–49.

[16] Craig Gentry, *Fully homomorphic encryption using ideal lattices* In STOC, ACM, 2009, pp. 169–178.

[17] ———, *Computing arbitrary functions of encrypted data*, Commun. ACM, **53**(3) (2010), 97–105.

[18] Craig Gentry, Shai Halevi, Mariana Raykova, and Daniel Wichs, *Outsourcing private ram computation*, IACR Crypt. ePrint Arch., 2014:148, 2014.

[19] Craig Gentry, Shai Halevi, and Nigel P. Smart, *Fully homomorphic encryption with polylog overhead*, In EUROCRYPT, Springer, 2012, pp. 465–482.

[20] Craig Gentry, Amit Sahai, and Brent Waters, *Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based*, In CRYPTO, Springer, 2013, pp. 75–92.

[21] Shafi Goldwasser and Silvio Micali, *Probabilistic encryption and how to play mental poker keeping secret all partial information*, In STOC, ACM, 1982, pp. 365–377.

[22] Shafi Goldwasser and Guy N. Rothblum, *On best-possible obfuscation*, In TCC, Springer, 2007, pp. 194–213.

[23] Russell Impagliazzo, *A personal view of average-case complexity*, In Structure in Complexity Theory Conference, IEEE, 1995, pp. 134–147.

[24] Joe Kilian, *Founding cryptography on oblivious transfer*, In STOC, ACM, 1988, pp. 20–31.

[25] Silvio Micali, 1988, Personal communication to Joe Kilian in [24].

[26] Daniele Micciancio and Petros Mol, *Pseudorandom knapsacks and the sample complexity of lwe search-to-decision reductions*, In CRYPTO, Springer, 2011, pp. 465–484.

[27] Daniele Micciancio and Chris Peikert, *Trapdoors for lattices: Simpler, tighter, faster, smaller* In EUROCRYPT, Springer, 2012, pp. 700–718.

[28] Koji Nuida, *A simple framework for noise-free construction of fully homomorphic encryption from a special class of non-commutative groups*, IACR Cryptology ePrint Archive, 2014:097, 2014.

[29] Chris Peikert, *Public-key cryptosystems from the worst-case shortest vector problem: extended abstract*, In STOC, ACM, 2009, pp. 333–342.

[30] Nicholas Pippenger and Michael J Fischer, *Relations among complexity measures*, Journal of the ACM (JACM) **26**(2) (1979), 361–381.

[31] Oded Regev, *On lattices, learning with errors, random linear codes, and cryptography*, In STOC, ACM, 2005, pp. 84–93.

[32] Ron Rivest, Leonard Adleman, and Michael Dertouzos, *On data banks and privacy homomorphisms*, In Found. of Sec. Comp., 1978, pp. 169–180.

[33] Ronald L. Rivest, Adi Shamir, and Leonard M. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Commun. ACM **21**(2) (1978), 120–126.

[34] Claus-Peter Schnorr, *A hierarchy of polynomial time lattice basis reduction algorithms* Theor. comp. sci. **53**(2) (1987), 201–224.

[35] Nigel P. Smart and Frederik Vercauteren, *Fully homomorphic simd operations*, Des. Codes Cryptography **71**(1) (2014), 57–81.

[36] Terence Tao, *Finite time blowup for an averaged three-dimensional navier-stokes equation*, arXiv:1402.0290, 2014.

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598
E-mail: cbgentry@us.ibm.com

Social choice, computational complexity, Gaussian geometry, and Boolean functions

Ryan O'Donnell

Abstract. We describe a web of connections between the following topics: the mathematical theory of voting and social choice; the computational complexity of the Maximum Cut problem; the Gaussian Isoperimetric Inequality and Borell's generalization thereof; the Hypercontractive Inequality of Bonami; and, the analysis of Boolean functions. A major theme is the technique of reducing inequalities about Gaussian functions to inequalities about Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$, and then using induction on n to further reduce to inequalities about functions $f : \{-1,1\} \to \{-1,1\}$. We especially highlight De, Mossel, and Neeman's recent use of this technique to prove the Majority Is Stablest Theorem and Borell's Isoperimetric Inequality simultaneously.

Mathematics Subject Classification (2010). Primary 68Q87; Secondary 94C10, 60G15.

Keywords. Social choice, analysis of Boolean functions, Majority Is Stablest, Max-Cut, computational complexity, Gaussian geometry, isoperimetry, hypercontractivity.

(This survey gives only a sketch of various results, and is slightly imprecise in places. For more details on the topics described herein, see [69].)

1. Social choice and Boolean functions

We begin by discussing a problem concerning voting. This will motivate for us certain definitions involving *Boolean functions*; i.e., functions $f : \{-1,1\}^n \to \{-1,1\}$ (or more generally, $f : \{-1,1\}^n \to \mathbb{R}$) whose domain consists of *n-bit strings*. Suppose we have an election with n voters and 2 candidates, named -1 and 1. A *voting rule* is simply any Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, mapping the voters' votes to the winner of the election. The *majority rule* $\mathrm{Maj}_n : \{-1,1\}^n \to \{-1,1\}$, defined (for n odd) by $\mathrm{Maj}_n(x) = \mathrm{sgn}(x_1 + x_2 + \cdots + x_n)$, is perhaps the most natural and mathematically elegant voting rule, but a variety of others are used in practice. Several countries (the US and the UK, for example) elect their head of state via a two-level (weighted-)majority scheme. Other countries, unfortunately, have been known to use a *dictator* rule: $f(x) = x_i$ for some dictator $i \in [n]$. The mathematical field of *social choice* is concerned with the properties of various voting rules; for a survey, see e.g. [18].

Let's now imagine a twist on the scenario: The n voters decide on their votes, $x = (x_1, \ldots, x_n) \in \{-1,1\}^n$. However, due to faulty voting machines, each vote is independently *misrecorded* with probability $\delta \in [0,1]$. We denote the resulting list of votes by $\boldsymbol{y} \in \{-1,1\}^n$, and call it a *noisy copy* of the original votes x. We now ask: *What is the*

Proceedings of the International Congress of Mathematicians, Seoul, 2014

probability that the noise affects the outcome of the election? How does this probability depend on the voting rule f? To answer this question we also need a probabilistic model for how the original votes are cast. We make the simplest possible assumption — that they are uniformly random, denoted $\boldsymbol{x} \sim \{-1,1\}^n$. In the social choice literature this is called the Impartial Culture Assumption [32]. Let's introduce some mathematical notation for our scenario, using the more convenient parameter $\rho = 1 - 2\delta \in [-1,1]$:

Definition 1.1. Given $x \in \{-1,1\}^n$ and $\rho \in [-1,1]$, we say that the random vector $\boldsymbol{y}$ is a *ρ-correlated copy* of x if each coordinate $\boldsymbol{y}_i$ is independently set to x_i with probability $\frac{1}{2}(1+\rho)$ and set to $-x_i$ with probability $\frac{1}{2}(1-\rho)$. (For the more common case of $\rho \geq 0$, this is equivalent to setting $\boldsymbol{y}_i = x_i$ with probability ρ and making $\boldsymbol{y}_i$ uniformly random with probability $1-\rho$.) When $\boldsymbol{x} \sim \{-1,1\}^n$ is uniformly random and $\boldsymbol{y}$ is a ρ-correlated copy of $\boldsymbol{x}$, we call $(\boldsymbol{x}, \boldsymbol{y})$ a *ρ-correlated random pair of strings*. Note that this is actually symmetric in $\boldsymbol{x}$ and $\boldsymbol{y}$; an alternative definition is that each pair $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \{-1,1\}^2$ is chosen independently with $\mathbf{E}[\boldsymbol{x}_i] = \mathbf{E}[\boldsymbol{y}_i] = 0$ and $\mathbf{E}[\boldsymbol{x}_i \boldsymbol{y}_i] = \rho$.

Definition 1.2. For $\rho \in [-1,1]$, the operator T_ρ acts on Boolean functions $f : \{-1,1\}^n \to \mathbb{R}$ via

$$\mathrm{T}_\rho f(x) = \mathop{\mathbf{E}}_{\boldsymbol{y} \text{ a } \rho\text{-correlated copy of } x}[f(\boldsymbol{y})].$$

We also define the *noise stability of f at ρ* to be

$$\mathbf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n}[f(\boldsymbol{x}) \cdot \mathrm{T}_\rho f(\boldsymbol{x})] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \ \rho\text{-correlated} \\ \text{strings}}}[f(\boldsymbol{x}) f(\boldsymbol{y})].$$

Note that in the special case $f : \{-1,1\}^n \to \{-1,1\}$,

$$\mathbf{Stab}_\rho[f] = 1 - 2 \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \ \rho\text{-correlated} \\ \text{strings}}}[f(\boldsymbol{x}) \neq f(\boldsymbol{y})].$$

Returning to the election scenario in which the voters' votes are misrecorded with probability δ, we see that the probability this affects the outcome of the election is precisely $\frac{1}{2} - \frac{1}{2}\mathbf{Stab}_{1-2\delta}[f]$. Thus the voting rules that minimize this probability are precisely those which maximize the noise stability $\mathbf{Stab}_{1-2\delta}[f]$.

Let's focus on the more natural case of $0 < \rho < 1$, i.e., $0 < \delta < \frac{1}{2}$. It's obvious that the Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ that maximize $\mathbf{Stab}_\rho[f]$ are precisely the two constant functions $f(x) = \pm 1$. These functions are highly unfair as voting rules, so it's natural to make an assumption that rules them out. One common such assumption is that f is *unbiased*, meaning $\mathbf{E}[f(\boldsymbol{x})] = 0$; in other words, the two outcomes ± 1 are equally likely when the voters vote uniformly at random. A stronger, but still very natural, assumption is that f is *odd*, meaning $f(-x) = -x$. In the social literature this is called *neutrality*, meaning that the voting rule is not affected by changing the names of the candidates.

We might now ask which *unbiased* functions $f : \{-1,1\}^n \to \{-1,1\}$ maximize $\mathbf{Stab}_\rho[f]$. This problem can be solved easily using *Fourier analysis of Boolean functions*, the basic facts of which we now recall:

Fact 1.3. *Any $f : \{-1,1\}^n \to \mathbb{R}$ can be uniquely expressed as a multilinear polynomial,*

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \prod_{i \in S} x_i.$$

This is called the Fourier expansion *of f, and the coefficients $\widehat{f}(S) \in \mathbb{R}$ are called the* Fourier coefficients *of f. We have* Parseval's formula*,*

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[f(\boldsymbol{x})g(\boldsymbol{x})] = \sum_{S\subseteq[n]} \widehat{f}(S)\widehat{g}(S).$$

In particular, if $f : \{-1,1\}^n \to \{-1,1\}$ then $\sum_S \widehat{f}(S)^2 = 1$.

Fact 1.4. *The Fourier expansion of $\mathrm{T}_\rho f$ is*

$$\mathrm{T}_\rho f(x) = \sum_{S\subseteq[n]} \rho^{|S|}\widehat{f}(S)\prod_{i\in S} x_i$$

and hence $\mathbf{Stab}_\rho[f] = \sum_S \rho^{|S|}\widehat{f}(S)^2$.

Using these facts, the following is an exercise:

Fact 1.5. *Assume $0 < \rho < 1$. Then $\mathbf{Stab}_\rho[f] \le \rho$ holds for all unbiased $f : \{-1,1\}^n \to \{-1,1\}$, with equality iff f is a (possibly negated) dictator function, $f(x) = \pm x_i$. Furthermore, $\mathbf{Stab}_{-\rho}[f] \ge -\rho$ holds for* all *$f : \{-1,1\}^n \to \{-1,1\}$, not necessarily unbiased, with the same equality conditions.*

This conclusion is somewhat disappointing from the standpoint of election fairness; it says that if our goal is to choose a voting rule that minimizes the effect of misrecorded votes (assuming $0 < \delta < \frac{1}{2}$), the "best" choice is dictatorship (or negated-dictatorship).

Incidentally, this is precisely the disappointment that occurs in *Arrow's Theorem* [6], the seminal result in social choice theory. In brief, Arrow's Theorem is concerned with what happens when n voters try to rank *three* candidates by means of holding three pairwise elections using Boolean voting rule f. The well-known *Condorcet Paradox* [22] is that for some f — including $f = \mathrm{Maj}_n$ — it is possible to get an "irrational" outcome in which the electorate prefers Candidate A to Candidate B, prefers Candidate B to Candidate C, and prefers Candidate C to Candidate A. Arrow showed that the only f's which *always* yield "rational" outcomes are dictators and negated-dictators. Kalai [46] gave a very elegant Fourier-analytic proof of Arrow's Theorem by noting that when the voters' individual rankings are uniformly random, the probability of a rational outcome is precisely $\frac{3}{4} - \frac{3}{4}\mathbf{Stab}_{-\frac{1}{3}}[f]$ (which also equals $\frac{3}{4} + \frac{3}{4}\mathbf{Stab}_{\frac{1}{3}}[f]$ for odd f). Then Arrow's conclusion follows from Fact 1.5. Kalai also obtained a robust version of Arrow's Theorem by using the *FKN Theorem* [31] from the analysis of Boolean functions: Any f that achieves a rational outcome with probability at least $1 - \delta$ must agree with some (negated-)dictator on all but an $O(\delta)$-fraction of inputs.

Just as we ruled out constant functions f by insisting on unbiasedness, we might also try to rule out dictatorships (and similar functions) by insisting that f give only negligible *influence* to each individual voter. Here we refer to the following definitions:

Definition 1.6. Let $f : \{-1,1\}^n \to \mathbb{R}$. For $i \in [n]$, the *(discrete) ith derivative* is

$$\mathrm{D}_i f(x) = \tfrac{f(x_1,\dots,x_{i-1},1,x_i,\dots,x_n) - f(x_1,\dots,x_{i-1},-1,x_i,\dots,x_n)}{2} = \sum_{S\ni i} \widehat{f}(S) \prod_{j\in S\setminus\{i\}} x_j.$$

The *ith influence of f* is

$$\mathbf{Inf}_i[f] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[\mathrm{D}_i f(\boldsymbol{x})^2] = \sum_{S\ni i} \widehat{f}(S)^2.$$

Note that when $f : \{-1,1\}^n \to \{-1,1\}$ we also have

$$\mathbf{Inf}_i[f] = \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \{-1,1\}^n}[f(\boldsymbol{x}) \neq f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{i-1}, -\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \dots, \boldsymbol{x}_n)].$$

If $f : \{-1,1\}^n \to \{-1,1\}$ is a voting rule, $\mathbf{Inf}_i[f]$ represents the probability that the ith voter's vote is pivotal for the outcome. (This notion was originally introduced by the geneticist Penrose [73]; it was independently popularized in the social choice literature by the lawyer Banzhaf [10].) The ith influence also has an interpretation in terms of the "geometry" of the discrete cube graph: if we think of $f : \{-1,1\}^n \to \{-1,1\}$ as the indicator of a vertex set $A \subseteq \{-1,1\}^n$, then $\mathbf{Inf}_i[f]$ is fraction of edges in the ith coordinate direction that are on A's boundary.

In the interest of fairness, one might want to disallow voting rules $f : \{-1,1\}^n \to \{-1,1\}$ that give unusually large influence to any one voter. This would disqualify a dictator voting rule like $f(x) = x_i$ since it has $\mathbf{Inf}_i[f] = 1$ (which is maximum possible). On the other hand, the majority voting rule is quite fair in this regard, since all of its influences quite small: using Stirling's formula one can compute $\mathbf{Inf}_i[\mathrm{Maj}_n] \sim \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}} \xrightarrow{n\to\infty} 0$ for all $i \in [n]$.

We can now ask a question that will occupy us for a significant portion of this survey:

Question 1.7. *Let $0 < \rho < 1$. Assume $f : \{-1,1\}^n \to \{-1,1\}$ is unbiased and satisfies $\max_i\{\mathbf{Inf}_i[f]\} \leq o_n(1)$. How large can $\mathbf{Stab}_\rho[f]$ be?*

We can think of this question as asking for the "fair" voting rule that minimizes the effect of misrecorded votes in a noisy election. Alternatively, the case of $\rho = \frac{1}{3}$ corresponds to asking for the "fair" odd voting rule which maximizes the probability of a "rational" outcome in the context of Arrow's Theorem.

Since majority rule seems like a fair voting scheme, it's natural to ask how well it does. For $n \to \infty$, this can be estimated using the Central Limit Theorem:

$$\begin{aligned}\mathbf{Stab}_\rho[\mathrm{Maj}_n] &= \mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \ \rho\text{-correlated} \\ \text{strings}}}\left[\mathrm{sgn}\left(\tfrac{\boldsymbol{x}_1+\cdots+\boldsymbol{x}_n}{\sqrt{n}}\right)\mathrm{sgn}\left(\tfrac{\boldsymbol{y}_1+\cdots+\boldsymbol{y}_n}{\sqrt{n}}\right)\right] \\ &\xrightarrow{n\to\infty} \mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}') \ \rho\text{-correlated} \\ \text{Gaussians}}}[\mathrm{sgn}(\boldsymbol{z})\mathrm{sgn}(\boldsymbol{z}')] = 1 - 2\,\mathbf{Pr}[\mathrm{sgn}(\boldsymbol{z}) \neq \mathrm{sgn}(\boldsymbol{z}')],\end{aligned}$$

where we say $(\boldsymbol{z}, \boldsymbol{z}')$ is a *ρ-correlated pair of Gaussians* if the random variables $\boldsymbol{z}, \boldsymbol{z}'$ are joint standard normals with $\mathbf{E}[\boldsymbol{z}\boldsymbol{z}'] = \rho$. An equivalent definition is that $\boldsymbol{z} = \langle \vec{u}, \vec{\boldsymbol{g}} \rangle$ and $\boldsymbol{z}' = \langle \vec{v}, \vec{\boldsymbol{g}} \rangle$, where $\vec{\boldsymbol{g}}$ is drawn from the standard d-dimensional Gaussian distribution γ_d and $\vec{u}, \vec{v} \in \mathbb{R}^d$ are any two unit vectors satisfying $\langle \vec{u}, \vec{v} \rangle = \rho$. (In particular, we can take $\boldsymbol{z} = \vec{\boldsymbol{g}}_1$, $\boldsymbol{z}' = \rho\vec{\boldsymbol{g}}_1 + \sqrt{1-\rho^2}\vec{\boldsymbol{g}}_2$.) Using this latter definition, it's not hard to verify the following old [78] fact:

Proposition 1.8 (Sheppard's Formula). *If $(\boldsymbol{z}, \boldsymbol{z}')$ are ρ-correlated Gaussians, $-1 \leq \rho \leq 1$, then $\mathbf{Pr}[\mathrm{sgn}(\boldsymbol{z}) \neq \mathrm{sgn}(\boldsymbol{z}')] = \frac{1}{\pi}\arccos\rho$.*

Taking care with the error term in the Central Limit Theorem, one may deduce:

Proposition 1.9. *For fixed $-1 < \rho < 1$,*

$$\mathbf{Stab}_\rho[\mathrm{Maj}_n] = 1 - \tfrac{2}{\pi}\arccos\rho + O(\tfrac{1}{\sqrt{n}}).$$

As a corollary, the probability of a "rational" outcome when using Maj_n in a three-way election tends to $\frac{3}{2\pi}\arccos(-\frac{1}{3}) \approx 91\%$, a fact known as *Guilbaud's Theorem* [38].

Is there a "fair" voting rule with even higher noise stability? In 2004, Khot et al. [51, 52] conjectured the result below, stating that majority essentially gives the best possible answer to Question 1.7. A year later their conjecture was proven by Mossel et al. [66, 67]:

Theorem 1.10 ("Majority Is Stablest Theorem"). *Fix* $0 < \rho < 1$. *Assume* $f : \{-1,1\}^n \to [-1,1]$ *satisfies* $\mathbf{E}[f(\boldsymbol{x})] = 0$ *and* $\max_i\{\mathbf{Inf}_i[f]\} \le \epsilon$. *Then*

$$\mathbf{Stab}_\rho[f] \le 1 - \tfrac{2}{\pi}\arccos\rho + o_\epsilon(1).$$

(Furthermore, for $-1 < \rho < 0$ *the inequality holds in reverse and the hypothesis* $\mathbf{E}[f(\boldsymbol{x})] = 0$ *is unnecessary.)*

Peculiarly, the motivation in Khot et al. [51] for conjecturing the above had nothing to do with social choice and voting. Instead, the conjecture was precisely what was needed to establish the computational complexity of finding approximately maximum *cuts* in graphs. We discuss this motivation next.

2. The computational complexity of Max-Cut

The *Max-Cut* problem is the following fundamental algorithmic task: Given as input is an undirected graph $G = (V, E)$. The goal is to find a partition $V = V^+ \cup V^-$ so as to maximize the fraction of *cut* edges. Here we say $e \in E$ is "cut" if it has one endpoint in each of $V^\pm$. We write $\mathrm{Opt}(G)$ to denote the value of the best possible solution; i.e., the maximum fraction of edges in G that can be cut. For example, $\mathrm{Opt}(G) = 1$ iff G is bipartite.

Unfortunately, the Max-Cut problem is known to be *NP-hard* [49]. This means that there is no efficient (i.e., $\mathrm{poly}(|V|)$-time) algorithm for determining $\mathrm{Opt}(G)$, assuming the well-believed P $\neq$ NP Conjecture. Under the closely related coNP $\neq$ NP Conjecture, we can also state this difficulty as follows: It is *not* true that whenever G is a graph satisfying $\mathrm{Opt}(G) \le \beta$, there is a short (i.e., $\mathrm{poly}(|V|)$-length) proof of the statement "$\mathrm{Opt}(G) \le \beta$".

Max-Cut is perhaps the simplest nontrivial *constraint satisfaction problem (CSP)*. Rather than formally defining this class of problems, we'll simply give two more examples. In the *Max-3Lin* problem, given is a system of equations over $\mathbb{F}_2$, each of the form "$x_{i_1} + x_{i_2} + x_{i_3} = b$"; the task is to find an assignment to the variables $x_1, \dots, x_n$ so as to maximize the fraction of satisfied equations. In the *Max-3Coloring* problem, given is an undirected graph; the task is to color the vertices using 3 colors so as to maximize the fraction of bichromatic edges.

For all of these CSPs the task of determining $\mathrm{Opt}(\cdot)$ is NP-hard. One way to cope with this difficulty is to seek *approximation algorithms*:

Definition 2.1. Let $0 \le \alpha \le \beta \le 1$. Algorithm $\mathcal{A}$ is said to be (α, β)-*approximating* for a certain CSP (e.g., Max-Cut) if it has the following guarantee: For every input G satisfying $\mathrm{Opt}(G) \ge \beta$, the algorithm finds a solution of value at least α. If $\mathcal{A}$ is a randomized algorithm, we allow it to achieve value at least α *in expectation*. Note that a fixed $\mathcal{A}$ may be (α, β)-approximating for many pairs (α, β) simultaneously.

Example 2.2. There is a simple greedy algorithm that is $(1,1)$-approximating for Max-Cut; i.e., given a bipartite G, it finds a bipartition. Similarly, one can efficiently $(1,1)$-approximate Max-3Lin using Gaussian elimination. On the other hand, $(1,1)$-approximating Max-3Coloring — i.e., validly 3-coloring 3-colorable graphs — is NP-hard. For Max-3Lin the near-trivial algorithm of outputting $x_1 = \cdots = x_n = B$, where B is the more common "right-hand side" of the system, is a $(\frac{1}{2}, \beta)$-approximation for every $\frac{1}{2} \leq \beta \leq 1$. One can also get an efficient $(\frac{1}{2}, \beta)$-approximation for Max-Cut (for any β) either by a simple greedy algorithm, or by outputting a *random* partition $V = V^+ \cup V^-$. The classical statement that "Max-Cut is NP-hard" is equivalent to stating that *there exists* $\frac{1}{2} < \beta < 1$ such that (β, β)-approximating Max-Cut is NP-hard (in fact, this is true for all $\frac{1}{2} < \beta < 1$).

In the case of Max-3Lin, it is a rather astonishing fact that the trivial approximation algorithms mentioned above are best possible assuming P $\neq$ NP; this is a celebrated result of Håstad [40, 41] combining "PCP technology" [4, 5, 13, 29] and Fourier analysis of Boolean functions:

Theorem 2.3. *For any $\delta > 0$, it's NP-hard to $(\frac{1}{2} + \delta, 1 - \delta)$-approximate Max-3Lin.*

For quite a long time, it was not known how to do any better even for the much simpler problem of Max-Cut. This changed in 1994 with the famous and sophisticated result of Goemans and Williamson [34, 35] (see also [23]):

Theorem 2.4. *There is an efficient algorithm that $(\frac{\theta}{\pi}, \frac{1}{2} - \frac{1}{2}\cos\theta)$-approximates Max-Cut for every $\theta \in [\theta_{\mathrm{GW}}, \pi]$, where $\theta_{\mathrm{GW}} \approx .74\pi$ is the positive solution of $\tan(\frac{\theta}{2}) = \theta$. E.g., the Goemans–Williamson algorithm simultaneously $(\frac{3}{4}, \frac{1}{2} + \frac{1}{2\sqrt{2}})$-approximates, $(\frac{4}{5}, \frac{5}{8} + \frac{\sqrt{5}}{8})$-approximates, and $(1 - \frac{2}{\pi}\sqrt{\epsilon} - o(\sqrt{\epsilon}), 1 - \epsilon)$-approximates Max-Cut.*

(Variants of the Goemans–Williamson algorithm that perform well for $\theta < \theta_{\mathrm{GW}}$ are also known.)

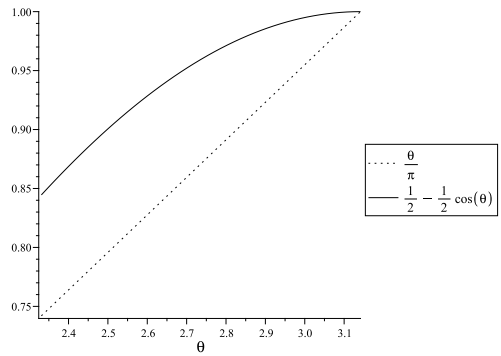


Figure 2.1.

Briefly, the algorithm works as follows: Given a graph $G = (V, E)$, one considers the following *semidefinite programming* optimization problem:

$$\begin{aligned} \mathrm{SDPOpt}(G) \;&=\; \max \quad \operatorname*{avg}_{(v,w)\in E} \left[\tfrac{1}{2} - \tfrac{1}{2}\langle \vec{U}(v), \vec{U}(w)\rangle\right] \\ &\text{subject to} \quad \vec{U} : V \to S^{d-1}. \end{aligned} \tag{SDP}$$

Here one also maximizes over all $d \in \mathbb{Z}^+$, although one can show that it suffices to take $d = |V|$. Essentially, the optimization problem (SDP) seeks to assign a unit vector to each vertex in V so that edges in G are spread as far apart as possible. It's easy to see that if d is fixed to 1 (so that $\vec{U} : V \to \{-1, 1\}$) then (SDP) is identical to the Max-Cut problem; therefore $\mathrm{Opt}(G) \leq \mathrm{SDPOpt}(G)$ always. Surprisingly, although computing $\mathrm{Opt}(G)$ is intractable, one can efficiently compute $\mathrm{SDPOpt}(G)$. (Roughly speaking, the reason is that if we introduce real variables $\rho_{vw} = \langle \vec{U}(v), \vec{U}(w) \rangle$, then (SDP) is equivalent to maximizing a linear function of the ρ_{vw}'s over an explicit convex subset of $\mathbb{R}^{|V| \times |V|}$, namely the set of all positive semidefinite matrices $R = (\rho_{vw})_{v,w \in V}$ with 1's on the diagonal.)

Thus (SDP) gives us an efficiently-computable upper bound on $\mathrm{Opt}(G)$. One may hope that it is a relatively "good" upper bound, and that furthermore one can prove this constructively by providing an efficient algorithm which converts the optimum "vector solution" $(\vec{U}^*(v))_{v \in V}$ to a good "± 1 solution" $(U^*(v))_{v \in V}$ — i.e., a good bipartition of V. Goemans and Williamson fulfilled this hope, as follows: Their algorithm first chooses $\vec{\boldsymbol{g}}$ to be a standard d-dimensional Gaussian and then it outputs the bipartition of G defined by $U^*(v) = \langle \vec{U}^*(v), \vec{\boldsymbol{g}} \rangle$. Using Sheppard's Formula, it's not hard to show that this establishes Theorem 2.4.

The Goemans–Williamson algorithm was originally considered to be quite complex for such a simple CSP as Max-Cut; furthermore, its approximation guarantee seemed quite peculiar. More than one paper [28, 30] suggested the research goal of improving this approximation guarantee. Furthermore, the best known NP-hardness result for the problem (from [40, 82]) does not match the algorithm. For example, it's known that $(.875 + \delta, .9)$-approximating Max-Cut is NP-hard for all $\delta > 0$, and the Goemans–Williamson algorithm achieves $(\alpha, .9)$-approximation for $\alpha = 1 - \frac{1}{\pi} \arccos \frac{4}{5} \approx .795$. But whether cutting 80% of the edges in a graph G with $\mathrm{Opt}(G) = 90\%$ is polynomial-time solvable or is NP-hard is unknown.

Nevertheless, in 2004 Khot et al. [51] obtained the following "surprising" [45] result: Under the *Unique Games Conjecture* [50] (a notorious conjecture in computational complexity not related to Max-Cut), the Majority Is Stablest Theorem implies that there is no efficient algorithm beating the Goemans–Williamson approximation guarantee (at least for $\theta \in [\theta_{\mathrm{GW}}, \pi]$; see [70] for optimal results when $\theta < \theta_{\mathrm{GW}}$). We remark that the while the Unique Games Conjecture is believable, its status is vastly more uncertain than the P $\neq$ NP conjecture).

Let us briefly explain what the Majority Is Stablest Theorem has to do with the complexity of the Max-Cut problem. As shown in [51], the advantage of the Unique Games Conjecture (as opposed to just the P $\neq$ NP assumption) is that it makes the "Håstad PCP technology" much easier to use. Very roughly speaking, it implies that to establish intractability of beating $(\frac{\theta}{\pi}, \frac{1}{2} - \frac{1}{2} \cos \theta)$-approximation, it suffices to find certain so-called "gadget graphs" for the Max-Cut problem. Precisely speaking, these gadget graphs need to have the following properties:

- The vertex set V should be $\{-1, 1\}^n$. (As a consequence, bipartitions of V correspond to Boolean functions $f : \{-1, 1\}^n \to \{-1, 1\}$.)
- The bipartitions given by the n "dictators" $f(x) = x_i$ should each cut at least a $\frac{1}{2} - \frac{1}{2} \cos \theta$ fraction of the edges.
- Any bipartition which is not "noticeably correlated" with a dictator partition should not cut "essentially more" than a $\frac{\theta}{\pi}$ fraction of the edges. More precisely, if f :

$\{-1,1\}^n \to \{-1,1\}$ is any bipartition of V with $\max_i\{\mathbf{Inf}_i[f]\} \le \epsilon$, then the fraction of edges it cuts is at most $\frac{\theta}{\pi} + o_\epsilon(1)$.

Actually, it's also acceptable for these gadgets to be *edge-weighted* graphs, with nonnegative edge-weights summing to 1. Khot et al. suggested using the *noisy hypercube* graph on vertex set $\{-1,1\}^n$, in which the weight on edge $(u,v) \in \{-1,1\}^n \times \{-1,1\}^n$ is precisely $\mathbf{Pr}[\boldsymbol{x} = u, \boldsymbol{y} = v]$ when $(\boldsymbol{x}, \boldsymbol{y})$ are a $(\cos\theta)$-correlated random strings (note that $\rho = \cos\theta < 0$ for $\theta \in [\theta_{\mathrm{GW}}, \pi]$). Such gadget graphs have the first two properties above, and the Majority Is Stablest Theorem precisely implies that they also have the third property. It's somewhat surprising that the technical properties required for this Unique Games/PCP-based hardness result correspond so perfectly to a natural problem about voting theory.

Thus subject to the Unique Games Conjecture, no efficient algorithm can improve on the Goemans–Williamson Max-Cut approximation guarantee. In particular, this means that there must be infinite families of graphs on which the Goemans–Williamson algorithm performs no better than the guarantee established in Theorem 2.4. As first shown by Karloff [48], the noisy hypercube graphs G also serve as examples here: Though they have $\mathrm{Opt}(G) = \frac{1}{2} - \frac{1}{2}\cos\theta$, one optimal solution of (SDP) for these graphs is $\vec{U}^*(v) = v/\sqrt{d}$, and applying the Goemans–Williamson algorithm to these vectors will indeed give a bipartition cutting only a $\frac{\theta}{\pi}$ fraction of edges in expectation.

Before turning our attention more fully to the Majority Is Stablest Theorem, we should mention a far-reaching generalization of the above-described work in complexity theory, namely the Raghavendra Theory of CSP approximation. Raghavendra [74] showed that for *all* CSPs (not just Max-Cut), the natural analogue of the Goemans–Williamson SDP algorithm has optimal approximation guarantee among all efficient algorithms, subject to the Unique Games Conjecture. This theory will be discussed further in our concluding Section 7.

3. Borell's isoperimetric inequality

The Majority Is Stablest Theorem concerns Boolean functions, but thanks to the Central Limit Theorem it includes as a "special case" a certain inequality concerning *Gaussian geometry* first proved by Borell [17]. (In this field, the idea that Boolean inequalities imply Gaussian inequalities dates back to the work of Gross [37] on the Log-Sobolev Inequality.) To state this Gaussian inequality we first make some definitions:

Definition 3.1. Given $z \in \mathbb{R}^d$ and $\rho \in [-1,1]$, we say that the random vector $\boldsymbol{z}'$ is a *ρ-correlated Gaussian copy* of z if $\boldsymbol{z}'$ has the distribution $\rho z + \sqrt{1-\rho^2}\boldsymbol{g}$, where $\boldsymbol{g}$ is a standard d-dimensional Gaussian random vector. When $\boldsymbol{z}$ is itself a standard d-dimensional Gaussian and $\boldsymbol{z}'$ is a ρ-correlated Gaussian copy, we call $(\boldsymbol{z}, \boldsymbol{z}')$ a *ρ-correlated d-dimensional Gaussian pair*. An equivalent definition is that each pair of random variables $(\boldsymbol{z}_i, \boldsymbol{z}'_i)$ is a ρ-correlated pair of Gaussians (as defined in Section 1) and the pairs are independent for $i \in [d]$. Note that $(\boldsymbol{z}, \boldsymbol{z}')$ has the same distribution as $(\boldsymbol{z}', \boldsymbol{z})$.

Remark 3.2. The distribution of a ρ-correlated d-dimensional Gaussian pair $(\boldsymbol{z}, \boldsymbol{z}')$ is also rotationally symmetric in $\mathbb{R}^d$. Note that for large d we'll have $\|\boldsymbol{z}\|, \|\boldsymbol{z}'\| \sim \sqrt{d}$ and $\langle \boldsymbol{z}, \boldsymbol{z}' \rangle \sim \rho d$. Thus an intuitive picture to keep in mind when d is large is that $(\boldsymbol{z}, \boldsymbol{z}')$ is roughly distributed as a uniformly random pair of vectors of length $\sqrt{d}$ and angle $\arccos\rho$.

Definition 3.3. The *Ornstein–Uhlenbeck semigroup* of operators is defined as follows: For $\rho \in [-1,1]$, the operator U_ρ acts on functions $f : \mathbb{R}^d \to \mathbb{R}$ by

$$\mathrm{U}_\rho f(z) = \mathop{\mathbf{E}}_{\boldsymbol{z}' \text{ a } \rho\text{-correlated Gaussian copy of } z}[f(\boldsymbol{z}')].$$

We also define the *Gaussian noise stability of f at ρ* to be

$$\mathbf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}') \ \rho\text{-correlated} \\ d\text{-dimensional Gaussian pair}}}[f(\boldsymbol{z})f(\boldsymbol{z}')].$$

We can now state the "Gaussian special case" of Majority Is Stablest:

Theorem 3.4. *Fix $0 < \rho < 1$. Assume $h : \mathbb{R}^d \to [-1,1]$ satisfies $\mathbf{E}_{\boldsymbol{z}\sim\gamma_d}[h(\boldsymbol{z})] = 0$. Then its Gaussian noise stability satisfies*

$$\mathbf{Stab}_\rho[h] \le 1 - \tfrac{2}{\pi}\arccos\rho.$$

(Furthermore, for $-1 < \rho < 0$ the inequality holds in reverse and the hypothesis $\mathbf{E}[h] = 0$ is unnecessary.)

To obtain Theorem 3.4 from the Majority Is Stablest Theorem (at least for "nice enough" h), we use the fact that Gaussian random variables can be "simulated" by sums of many independent ± 1 random bits. More precisely, we can apply Majority Is Stablest to $f : \{-1,1\}^{dn} \to [-1,1]$ defined by

$$f(x_{1,1},\dots,x_{d,n}) = h\left(\tfrac{x_{1,1}+\cdots+x_{1,n}}{\sqrt{n}},\dots,\tfrac{x_{d,1}+\cdots+x_{d,n}}{\sqrt{n}}\right)$$

and then take $n \to \infty$ and use a d-dimensional Central Limit Theorem. (The assumption and error dependence on the influence bound ϵ disappears, because we have $\epsilon \to 0$ as $n \to \infty$.) Note that in Section 1 we saw exactly this limiting procedure in the case of $h = \mathrm{sgn} : \mathbb{R}^1 \to \{-1,1\}$ when we computed the limiting (Boolean) noise stability of Maj_n.

Theorem 3.4 was first proved by Borell in 1985 [17]. (In fact, Borell proved significant generalizations of the theorem, as discussed below.) In 2005, Mossel et al. [67] used it to prove the Majority Is Stablest Theorem by reducing the Boolean setting to the Gaussian setting. The key technical tool here was a "nonlinear" version of the Central Limit Theorem called the *Invariance Principle* (see also [76]). Briefly, the Invariance Principle implies that if $f : \{-1,1\}^n \to \mathbb{R}$ is a low-degree multilinear polynomial with small influences then the distributions of $f(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$ and $f(\boldsymbol{g}_1,\dots,\boldsymbol{g}_n)$ are "close", where $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ are independent ± 1 random variables and $\boldsymbol{g}_1,\dots,\boldsymbol{g}_n$ are independent Gaussians. The Invariance Principle has had many applications (e.g., in combinatorics [24], learning theory [47], pseudorandomness [62], social choice [64], sublinear algorithms [14], and the Raghavendra Theory of CSPs mentioned at the end of Section 2) but we won't discuss it further here. Instead, we'll outline in Section 6 an alternative, "purely discrete" proof of the Majority Is Stablest Theorem due to De, Mossel, and Neeman [20].

Let's now look more carefully at the geometric content of Theorem 3.4. Suppose $A \subset \mathbb{R}^d$ is a set with *Gaussian volume* $\gamma_d(A) = \frac{1}{2}$. Applying Theorem 3.4 with $h = 1 - 2\cdot 1_A$, and also writing $\theta = \arccos\rho \in (0, \frac{\pi}{2})$, one obtains the following:

Corollary 3.5. *For $0 \le \theta \le \frac{\pi}{2}$ and $A \subseteq \mathbb{R}^d$, define the* rotation sensitivity

$$\mathbf{RS}_A(\theta) = \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z},\boldsymbol{z}') \cos\theta\text{-correlated} \\ d\text{-dimensional Gaussian pair}}} [1_A(\boldsymbol{z}) \neq 1_A(\boldsymbol{z}')].$$

Then if $\gamma_d(A) = \frac{1}{2}$, we have $\mathbf{RS}_A(\theta) \ge \frac{\theta}{\pi}$.

By Sheppard's Formula, equality is obtained if $d = 1$ and $A = (-\infty, 0]$. In fact, by rotational symmetry of correlated Gaussians, equality is obtained when A is any halfspace through the origin in $\mathbb{R}^d$. (Geometrically, it's natural to guess that halfspaces minimize $\mathbf{RS}_A(\theta)$ among sets A of fixed Gaussian volume, using the intuition from Remark 3.2.) As shown in [55], this corollary is quite easy to prove for "many" values of θ:

Proof of Corollary 3.5 for $\theta = \frac{\pi}{2\ell}$, $\ell \in \mathbb{Z}^+$. Let $\boldsymbol{g}, \boldsymbol{g}'$ be independent d-dimensional Gaussians and define $\boldsymbol{z}^{(j)} = \cos(j\theta)\boldsymbol{g} + \sin(j\theta)\boldsymbol{g}'$ for $0 \le j \le \ell$. Then it's easy to check that $(\boldsymbol{z}^{(i)}, \boldsymbol{z}^{(j)})$ is a $\cos((j-i)\theta)$-correlated Gaussian pair. In particular, $\boldsymbol{z}^{(0)}$ and $\boldsymbol{z}^{(\ell)}$ are independent. Now using $\gamma_d(A) = \frac{1}{2}$ and a union bound we get

$$\tfrac{1}{2} = \mathbf{Pr}[1_A(\boldsymbol{z}^{(0)}) \neq 1_A(\boldsymbol{z}^{(\ell)})] \le \sum_{j=1}^{\ell} \mathbf{Pr}[1_A(\boldsymbol{z}^{(j-1)}) \neq 1_A(\boldsymbol{z}^{(j)})] = \ell \cdot \mathbf{RS}_A(\theta),$$

which is the desired inequality. □

Returning to Theorem 3.4, it states that if $(\boldsymbol{z}, \boldsymbol{z}')$ are ρ-correlated d-dimensional Gaussians ($0 < \rho < 1$) then halfspaces are the volume-$\frac{1}{2}$ sets which maximize $\mathbf{Pr}[\boldsymbol{z}, \boldsymbol{z}' \in A]$. In fact, halfspaces are also the optimizers at *any* fixed volume. Furthermore, if we generalize by looking for sets A, B of fixed volume maximizing $\mathbf{Pr}[\boldsymbol{z} \in A, \boldsymbol{z}' \in B]$, parallel halfspaces are again best. These isoperimetric facts (and more) were all originally proved by Borell [17]:

Theorem 3.6 ("Borell Isoperimetric Inequality"). *Fix $0 < \rho < 1$ and $0 \le \alpha, \beta \le 1$. Suppose $A, B \subseteq \mathbb{R}^d$ satisfy $\gamma_d(A) = \alpha$, $\gamma_d(B) = \beta$. Then if $(\boldsymbol{z}, \boldsymbol{z}')$ is a ρ-correlated d-dimensional Gaussian pair,*

$$\mathbf{Pr}[\boldsymbol{z} \in A, \boldsymbol{z}' \in B] \le \mathbf{Pr}[\boldsymbol{z} \in H, \boldsymbol{z}' \in H']$$

where H and H' are (any) parallel halfspaces satisfying $\gamma_d(H) = \alpha$, $\gamma_d(H') = \beta$. (If $-1 < \rho < 0$ then the inequality is reversed.) By rotational symmetry we may assume $H = (-\infty, \Phi^{-1}(\alpha)]$, $H' = (-\infty, \Phi^{-1}(\beta)] \subseteq \mathbb{R}$ and thus write the above as

$$\mathbf{Pr}[\boldsymbol{z} \in A, \boldsymbol{z}' \in B] \le \Lambda_\rho(\alpha, \beta) := \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{w},\boldsymbol{w}') \ \rho\text{-correlated} \\ \text{Gaussians}}} [\boldsymbol{w} \le \Phi^{-1}(\alpha), \boldsymbol{w}' \le \Phi^{-1}(\beta)].$$

In case $\alpha = \beta = \frac{1}{2}$, Sheppard's Formula implies

$$\mathbf{Pr}[\boldsymbol{z} \in A, \boldsymbol{z}' \in B] \le \Lambda_\rho(\tfrac{1}{2}, \tfrac{1}{2}) = \tfrac{1}{2} - \tfrac{1}{2\pi} \arccos \rho.$$

Borell's original proof of this theorem used the Gaussian symmetrization method due to Ehrhard [25] and was quite technical. Four other proofs are known. Beckner [12] pointed out

that the analogous isoperimetric inequality on the sphere is easy to prove by two-point symmetrization [7], and the Gaussian result can then be deduced via "Poincaré's limit" (see [19]). Mossel and Neeman [65] recently gave a slick proof using semigroup methods, and together with De [20] they gave another proof via Boolean functions. Finally, Eldan [27] gave the most recent new proof, using stochastic calculus.

We will describe De, Mossel, and Neeman's Boolean proof of Borell's Isoperimetric Inequality in Section 6. It has the advantage that it can be used to prove the Majority Is Stablest Theorem "at the same time" (using a few technical tricks from the original Invariance Principle-based proof, including *hypercontractivity*). But first, we'll spend some time discussing further special cases of Borell's Isoperimetric Inequality.

4. Hypercontractivity

Borell's Isoperimetric Inequality is very precise, giving the exact maximal value of $\mathbf{Pr}[\boldsymbol{z} \in A, \boldsymbol{z}' \in B]$ (when $(\boldsymbol{z}, \boldsymbol{z}')$ are ρ-correlated) for any fixed Gaussian volumes $\gamma_d(A) = \alpha$, $\gamma_d(B) = \beta$. A small downside is that this maximum value, $\Lambda_\rho(\alpha, \beta)$, does not have a nice closed-form expression except when $\alpha = \beta = \frac{1}{2}$. In the interesting regime of $\alpha, \beta \to 0$, however, we can get a closed form for its asymptotics. Let's do a rough "heuristic" estimation.

Suppose H, H' are parallel halfspaces of "small" Gaussian volume α, β, with $\alpha \leq \beta$. By rotational symmetry we can assume $H = [a, \infty), H' = [b, \infty) \subset \mathbb{R}$ for some "large" values $a \geq b > 0$. Precisely, we have $a = -\Phi^{-1}(\alpha)$, but speaking roughly we'll express this as $\alpha \approx \exp(-\frac{a^2}{2})$, as this is asymptotically correct up to lower-order factors. Similarly we'll write $\beta \approx \exp(-\frac{b^2}{2})$. We are interested in estimating $\mathbf{Pr}[\boldsymbol{g} \in H, \boldsymbol{g}' \in H']$, where $(\boldsymbol{g}, \boldsymbol{g}')$ are a ρ-correlated Gaussian pair. We'll actually take $\boldsymbol{g}' = \rho \boldsymbol{g} + \sqrt{1 - \rho^2}\boldsymbol{h}$, where $\boldsymbol{h}$ is a standard Gaussian independent of $\boldsymbol{g}$. To start the estimation, by definition we have $\mathbf{Pr}[\boldsymbol{g} \in H] \approx \exp(-\frac{a^2}{2})$. Further, conditioned on $\boldsymbol{g} \in H$ we will almost surely have that $\boldsymbol{g}$ is only "barely" larger than a. Thus we expect $\boldsymbol{g}'$ to be conditionally distributed roughly as $\rho a + \sqrt{1 - \rho^2}\boldsymbol{h}$. In this case, $\boldsymbol{g}'$ will be in H' if and only if $\boldsymbol{h} \geq (b - \rho a)/\sqrt{1 - \rho^2}$. Under the assumption that $b - \rho a \geq 0$, the probability of this is, roughly again, $\exp(-\frac{(b-\rho a)^2}{2(1-\rho^2)})$. All in all, these calculations "suggest" that

$$\Lambda_\rho(\alpha, \beta) = \mathbf{Pr}[\boldsymbol{g} \in H, \boldsymbol{g}' \in H'] \approx \exp(-\tfrac{a^2}{2}) \exp(-\tfrac{(b-\rho a)^2}{2(1-\rho^2)}) = \exp\left(-\tfrac{1}{2}\tfrac{a^2 - 2\rho ab + b^2}{1-\rho^2}\right)$$

(under the assumption that $\alpha \approx \exp(-\frac{a^2}{2}) \leq \exp(-\frac{b^2}{2}) \approx \beta$ are "small", with $b \geq \rho a$). Since Borell's Isoperimetric Inequality tells us that parallel halfspaces are maximizers, we might optimistically guess the following:

Theorem 4.1 ("Gaussian Small-Set Expansion Theorem"). *Let $0 < \rho < 1$. Let $A, B \subseteq \mathbb{R}^d$ have Gaussian volumes $\exp(-\frac{a^2}{2}), \exp(-\frac{b^2}{2})$, respectively, and assume $0 \leq \rho a \leq b \leq a$. Then*

$$\mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z}, \boldsymbol{z}') \ \rho\text{-correlated} \\ d\text{-dimensional Gaussian pair}}} [\boldsymbol{z} \in A, \boldsymbol{z}' \in B] \leq \exp\left(-\tfrac{1}{2}\tfrac{a^2 - 2\rho ab + b^2}{1-\rho^2}\right).$$

In particular, if $A \subseteq \mathbb{R}^d$ has $\gamma_d(A) = \alpha$ then

$$\mathbf{Stab}_\rho[1_A] \le \alpha^{\frac{2}{1+\rho}} \iff \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z},\boldsymbol{z}') \ \rho\text{-correlated} \\ d\text{-dimensional Gaussian pair}}} [\boldsymbol{z}' \in A \mid \boldsymbol{z} \in A] \le \alpha^{\frac{1-\rho}{1+\rho}}. \tag{4.1}$$

Indeed this theorem is correct, and it can be formally deduced from Borell's Isoperimetric Inequality. We'll outline a more direct proof shortly, but first let's discuss its content. The one-set statement (4.1) says that if A is any "small" subset of Gaussian space (think of α as tending to 0) and ρ is bounded away from 1 (say $\rho = 1-\delta$), then a ρ-noisy copy of a random point in A will almost certainly (i.e., except with probability $\alpha^{\delta/(2+\delta)}$) be outside A.

One might ask whether a similar statement is true for subsets of the discrete cube $\{-1,1\}^n$. As we saw with Majority Is Stablest implying Theorem 3.4, isoperimetric inequalities on the discrete cube typically imply the analogous statement in Gaussian space, by the Central Limit Theorem. On the other hand, the converse does not generally hold; this is because there are subsets of $\{-1,1\}^n$ like the dictators $\{x : x_i = 1\}$, or more generally "subcubes" $\{x : x_{i_1} = \cdots = x_{i_k} = 1\}$, which have no analogue in Gaussian space. In particular, one has to rule out dictators using the "small-influences" condition in order for the Boolean analogue of Borell's theorem, namely the Majority Is Stablest Theorem, to be true. However it *is* often true that asymptotic isoperimetric inequalities for "small" subsets of Gaussian space also hold in the Boolean setting with no influences assumption; this is because *small* subcubes and *small* Hamming balls (the Boolean analogue of Gaussian halfspaces) have similar isoperimetric properties in $\{-1,1\}^n$. In particular, it turns out that Theorem 4.1 holds identically in $\{-1,1\}^n$:

Theorem 4.2 ("Boolean Small-Set Expansion Theorem"). *Let $0 < \rho < 1$. Let $A, B \subseteq \{-1,1\}^n$ have volumes $\frac{|A|}{2^n} = \exp(-\frac{a^2}{2})$, $\frac{|B|}{2^n} = \exp(-\frac{b^2}{2})$, and assume $0 \le \rho a \le b \le a$. Then*

$$\mathop{\mathbf{Pr}}_{(\boldsymbol{x},\boldsymbol{y}) \ \rho\text{-correlated strings}} [\boldsymbol{x} \in A, \boldsymbol{y} \in B] \le \exp\left(-\tfrac{1}{2}\tfrac{a^2 - 2\rho ab + b^2}{1-\rho^2}\right).$$

In particular, if $\frac{|A|}{2^n} = \alpha$ then

$$\mathbf{Stab}_\rho[1_A] \le \alpha^{\frac{2}{1+\rho}} \iff \mathop{\mathbf{Pr}}_{(\boldsymbol{x},\boldsymbol{y}) \ \rho\text{-correlated strings}} [\boldsymbol{x} \in A \mid \boldsymbol{y} \in A] \le \alpha^{\frac{1-\rho}{1+\rho}}. \tag{4.2}$$

This theorem is formally stronger than its Gaussian counterpart Theorem 4.1, by virtue of the Central Limit Theorem. In fact, there is a related *functional inequality* which is even stronger; this is the crucial *Hypercontractive Inequality* first proved by Bonami [15].

Theorem 4.3 ("Boolean Hypercontractive Inequality"). *Let $f, g : \{-1,1\}^n \to \mathbb{R}$, let $r, s \ge 0$, and assume $0 \le \rho \le \sqrt{rs} \le 1$. Then*

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{y}) \ \rho\text{-correlated}} [f(\boldsymbol{x})g(\boldsymbol{y})] \le \|f\|_{1+r}\|g\|_{1+s}.$$

(Here we are using L^p-norm notation, $\|f\|_p = \mathbf{E}_{\boldsymbol{x} \sim \{-1,1\}^n}[|f(\boldsymbol{x})|^p]^{1/p}$.)

To recover Theorem 4.2, one simply applies the Hypercontractive Inequality with $f = 1_A$, $g = 1_B$ and optimizes the choice of r, s. (We mention that this deduction was first noted, in its "reverse" form, by Mossel et al. [68].) The Gaussian analogue of the Boolean Hypercontractive Inequality also holds; indeed, the traditional proof of it (say, in [44]) involves first proving the Boolean inequality and then applying the Central Limit Theorem.

Another interpretation of the Hypercontractive Inequality is as a "generalized Hölder's inequality". In fact, its $\rho = 1$ case (corresponding to $\boldsymbol{y} \equiv \boldsymbol{x}$) is *identical* to Hölder's inequality (since the hypothesis $\sqrt{rs} = 1$ is identical to $(1+s)' = 1+r$). The Hypercontractive Inequality shows that as $\boldsymbol{x}$ and $\boldsymbol{y}$ become less and less correlated, one can put smaller and smaller norms of f and g on the right-hand side. (In the ultimate case of $\rho = 0$, meaning $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, one gets the trivial inequality $\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{y})] \leq \|f\|_1\|g\|_1$.)

Speaking of Hölder's inequality, we should mention that it can be used to show that Theorem 4.3 is equivalent to the following more traditional formulation of the Hypercontractive Inequality: For $f : \{-1,1\}^n \to \mathbb{R}$, $1 \leq p \leq q \leq \infty$:

$$\|\mathrm{T}_\rho f\|_q \leq \|f\|_p \tag{4.3}$$

provided $0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$.

Writing $p = 1 + r$, $q = 1 + 1/s$, one uses the fact that $\|\mathrm{T}_\rho f\|_q = \sup\{\mathbf{E}[g \cdot \mathrm{T}_\rho f] : \|g\|_{q'} = 1\}$, and that the quantity inside the sup is the same as the left-hand side in Theorem 4.3. Here we see an explanation for the name of the inequality — it shows that T_ρ is not just a contraction in L^p but in fact is a "hypercontraction" from L^p to L^q. In this formulation, the inequality can be viewed as quantifying the "smoothing" effect of the T_ρ operator. By virtue of Fact 1.4 one can use this formulation to show that low-degree polynomials of independent ± 1 random variables are "reasonable", in the sense that their high norms are comparable to their 2-norm. However we won't pursue this interpretation any further here.

A wonderful fact about the Boolean Hypercontractive Inequality is that the $n = 1$ case implies the general n case by induction. Indeed, for the two-function form given in Theorem 4.3, the induction is almost trivial. If $(\boldsymbol{x}, \boldsymbol{y})$ are ρ-correlated and we write $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}')$ for $\boldsymbol{x}' \in \{-1,1\}^{n-1}$ (and similarly for $\boldsymbol{y}$), then

$$\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{y})] = \mathop{\mathbf{E}}_{(\boldsymbol{x}_1,\boldsymbol{y}_1)} \mathop{\mathbf{E}}_{(\boldsymbol{x}',\boldsymbol{y}')}[f_{\boldsymbol{x}_1}(\boldsymbol{x}')g_{\boldsymbol{y}_1}(\boldsymbol{y}')] \leq \mathop{\mathbf{E}}_{(\boldsymbol{x}_1,\boldsymbol{y}_1)}[\|f_{\boldsymbol{x}_1}\|_{1+r}\|g_{\boldsymbol{y}_1}\|_{1+s}],$$

by induction, where f_{x_1} denotes the restriction of f gotten by fixing the first coordinate to be x_1 (and similarly for g_{y_1}). Then defining the 1-bit functions $F(x_1) = \|f_{x_1}\|_{1+r}$ and $G(y_1) = \|g_{y_1}\|_{1+s}$ we have

$$\mathop{\mathbf{E}}_{(\boldsymbol{x}_1,\boldsymbol{y}_1)}[\|f_{\boldsymbol{x}_1}\|_{1+r}\|g_{\boldsymbol{y}_1}\|_{1+s}] = \mathop{\mathbf{E}}_{(\boldsymbol{x}_1,\boldsymbol{y}_1)}[F(\boldsymbol{x}_1)G(\boldsymbol{y}_1)] \leq \|F\|_{1+r}\|G\|_{1+s} = \|f\|_{1+r}\|g\|_{1+s},$$

where we used the $n = 1$ case of the Hypercontractive Inequality.

Thus to prove all of the Boolean and Gaussian Hypercontractivity and Small-Set Expansion theorems, it suffices to prove the $n = 1$ case of the Boolean Hypercontractive Inequality. In fact, by the Hölder trick we just need to prove (4.3) in the case $n = 1$. It's also easy to show that we can assume $f : \{-1,1\} \to \mathbb{R}$ is nonnegative, and by homogeneity we can also assume f has mean 1. Thus everything boils down to proving the following: If $0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}} \leq 1$ and $0 \leq \delta \leq 1$ then

$$\left(\tfrac{1}{2}(1+\rho\delta)^q + \tfrac{1}{2}(1-\rho\delta)^q\right)^{1/q} \leq \left(\tfrac{1}{2}(1+\delta)^p + \tfrac{1}{2}(1-\delta)^p\right)^{1/p}. \tag{4.4}$$

Note that if we think of δ as very small and perform a Taylor expansion, the above becomes

$$1 + \tfrac{1}{2}\rho^2(q-1)\delta^2 + \cdots \leq 1 + \tfrac{1}{2}(p-1)\delta^2 + \cdots.$$

This shows that the $\rho \le \sqrt{\frac{p-1}{q-1}}$ condition is necessary, and also that it's "essentially" sufficient assuming δ is small. However, we need to actually verify (4.4) for all $0 \le \delta \le 1$. For some simple values of p and q, this is easy. For example, if $p = 2$ and $q = 4$, establishing (4.4) amounts to noting that $1 + 2\delta^2 + \frac{1}{9}\delta^4 \le 1 + 2\delta^2 + \delta^4$. This is already enough to prove, say, the Boolean Small-Set Expansion statement (4.2) with parameter $\rho = \frac{1}{3}$. On the other hand, establishing (4.4) for all p, q and all δ is a little bit painful (albeit elementary). In the next section, we'll see a similar problem where this pain can be circumvented.

5. Bobkov's inequality and Gaussian isoperimetry

Let's now look at a different special case of Borell's Isoperimetric Inequality, namely the case where $B = A$ and $\rho \to 1^-$. Using the rotation sensitivity definition from Corollary 3.5, Borell's inequality tells us that if $A \subseteq \mathbb{R}^d$, and $H \subseteq \mathbb{R}^d$ is a halfspace of the same Gaussian volume, then $\mathbf{RS}_A(\delta) \ge \mathbf{RS}_H(\delta)$. Since we also have $\mathbf{RS}_A(0) = \mathbf{RS}_H(0) = 0$, it follows that $\mathbf{RS}'_A(0^+) \ge \mathbf{RS}'_H(0^+)$. (It can be shown that this derivative $\mathbf{RS}'_A(0^+)$ is always well-defined, though it may be ∞.) As we'll explain shortly, the derivative $\mathbf{RS}'_A(0^+)$ has a very simple meaning; up to a factor of $\sqrt{\frac{\pi}{2}}$, it is the *Gaussian surface area* of the set A. Thus Borell's Isoperimetric Inequality implies the following well-known result:

Theorem 5.1 ("Gaussian Isoperimetric Inequality"). *Let $A \subseteq \mathbb{R}^d$ have Gaussian volume $\gamma_d(A) = \alpha$, and let $H \subseteq \mathbb{R}^d$ be any halfspace with $\gamma_d(H) = \alpha$. Then*

$$\gamma_d^+(A) \ge \gamma_d^+(H). \tag{5.1}$$

Here we are using the following definition:

Definition 5.2. The *Gaussian surface area* of $A \subseteq \mathbb{R}^d$ is

$$\gamma_d^+(A) = \sqrt{\frac{\pi}{2}} \cdot \mathbf{RS}'_A(0^+) = \lim_{\delta \to 0^+} \frac{\gamma_d((\partial A)^{+\delta/2})}{\delta} = \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \gamma_d}[\|\nabla 1_A(\boldsymbol{z})\|] = \int_{\partial A} \varphi(x)\, dx.$$

The first equation may be taken as the definition, and the remaining equations hold assuming A is "nice enough" (for technical details, see [1, 2, 2, 3, 43, 63]).

To get a feel for the definition, let's "heuristically justify" the second equality above, which relates the derivative of rotation sensitivity to the more natural-looking Gaussian Minkowski content of ∂A. We can think of

$$\mathbf{RS}'_A(0^+) = \frac{\mathbf{RS}_A(\delta)}{\delta} = \frac{1}{\delta} \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z}, \boldsymbol{z}') \text{ } \cos\delta\text{-correlated} \\ d\text{-dimensional Gaussian pair}}} [1_A(\boldsymbol{z}) \ne 1_A(\boldsymbol{z}')] \tag{5.2}$$

for "infinitesimal" δ. The last expression here can be thought of as the probability that the line segment $\boldsymbol{\ell}$ joining $\boldsymbol{z}, \boldsymbol{z}'$ crosses ∂A. Now for infinitesimal δ we have $\cos\delta \approx 1$ and $\sin\delta \approx \delta$; thus the distribution of $(\boldsymbol{z}, \boldsymbol{z}')$ is essentially that of $(\boldsymbol{g}, \boldsymbol{g} + \delta\boldsymbol{g}')$ for $\boldsymbol{g}$, $\boldsymbol{g}'$ independent d-dimensional Gaussians. When $\boldsymbol{g}$ lands near ∂A, the length of the segment $\boldsymbol{\ell}$ in the direction of the nearby unit normal $\boldsymbol{v}$ to ∂A will have expectation $\mathbf{E}[|\langle \delta\boldsymbol{g}', \boldsymbol{v}\rangle|] = \delta\,\mathbf{E}[|\mathrm{N}(0,1)|] = \sqrt{2/\pi}\cdot\delta$. Thus (5.2) should essentially be $\sqrt{2/\pi}\cdot\delta\cdot\gamma_d(\{z : \mathrm{dist}(z, \partial A) < \delta\})$, completing the heuristic justification of the second inequality in Definition 5.2.

Incidentally, it's easy to see that the Gaussian surface area of the one-dimensional halfspace $(-\infty, a] \subseteq \mathbb{R}$ is $\varphi(a)$; thus we can give an explicit formula for the right-hand side of (5.1):

Fact 5.3. *The right-hand side of* (5.1) *is the* Gaussian isoperimetric function*,*

$$\mathcal{I}(\alpha) = \varphi \circ \Phi^{-1}(\alpha) \in [0, \tfrac{1}{\sqrt{2\pi}}].$$

A remark: One easily checks that it satisfies the differential equation $\mathcal{I}\mathcal{I}'' + 1 = 0$, with boundary conditions $\mathcal{I}(0) = \mathcal{I}(1) = 0$.

The Gaussian Isoperimetric Inequality was originally independently proven by Borell [16] and by Sudakov and Tsirel'son [81]. Both proofs deduced it via Poincaré's limit from Lévy's Spherical Isoperimetric Inequality [61, 77]. (This is the statement that the fixed-volume subsets of a sphere's surface which minimize perimeter are caps — i.e., intersections of the sphere with a halfspace.) Ehrhard [25] subsequently developed his Gaussian symmetrization method to give a different proof. In 1997, Bobkov gave a surprising new proof by the same technique we saw in the last section: establishing a functional Boolean analogue by induction. We'll now outline this proof.

We start with the following equivalent functional form of the Gaussian Isoperimetric Inequality (first noted by Ehrhard [26]): For locally Lipschitz $f : \mathbb{R}^d \to [0, 1]$,

$$\mathcal{I}(\mathbf{E}[f(\boldsymbol{z})]) \leq \mathbf{E}[\|(\nabla f(\boldsymbol{z}), \mathcal{I}(f(\boldsymbol{z})))\|_2], \tag{5.3}$$

where $\boldsymbol{z} \sim \gamma_d$ and $\|\cdot\|_2$ denotes the usual Euclidean norm in $d+1$ dimensions. The Gaussian Isoperimetric Inequality for A can be deduced by taking $f = 1_A$; conversely, inequality 5.3 can be deduced from the Gaussian Isoperimetric Inequality by taking $A = \{(x, a) : f(x) \geq \Phi(a)\} \subseteq \mathbb{R}^{d+1}$. In turn, Bobkov showed that the above inequality can be deduced (by the usual Central Limit Theorem argument) from the analogous Boolean inequality:

Theorem 5.4 ("Bobkov's Inequality"). *For any $f : \{-1, 1\}^n \to [0, 1]$,*

$$\mathcal{I}(\mathbf{E}[f]) \leq \mathbf{E}[\|(\nabla f, \mathcal{I}(f))\|_2].$$

Here the expectation is with respect to the uniform distribution on $\{-1, 1\}^n$, and $\nabla f = (\mathrm{D}_1 f, \ldots, \mathrm{D}_n f)$.

Just as with the Hypercontractive Inequality, this inequality has the property that the $n = 1$ case implies the general n case by a fairly easy induction. Indeed, this induction uses no special property of $\mathcal{I}$ or the 2-norm:

Fact 5.5. *Let $J : [0, 1] \to \mathbb{R}^{\geq 0}$, and let $\|\cdot\|$ denote a fixed L^p-norm. Consider, for $f : \{-1, 1\}^n \to [0, 1]$, the following inequality:*

$$J(\mathbf{E}[f]) \leq \mathbf{E}[\|(\nabla f, J(f))\|]. \tag{5.4}$$

If this inequality holds for $n = 1$ then it holds for general n.

Now given a norm $\|\cdot\|$ we can seek the "largest" function J for which (5.4) holds when $n = 1$. As an aside, for the 1-norm $\|\cdot\|_1$ we may take $J(\alpha) = \alpha \log_2(1/\alpha)$, and this yields a form of the classic Edge Isoperimetric Inequality for the discrete cube [39], sharp for all

$\alpha = 2^{-k}, k \in \mathbb{Z}^+$. Returning to Bobkov's Inequality, the $n = 1$ case we need to verify is that

$$J(\alpha) \le \tfrac{1}{2}\sqrt{\delta^2 + J(\alpha+\delta)^2} + \tfrac{1}{2}\sqrt{\delta^2 + J(\alpha-\delta)^2} \tag{5.5}$$

when $J = \mathcal{I}$ and $\alpha \pm \delta \in [0,1]$. Bobkov used some (elementary) labor to show that this inequality indeed holds when $J = \mathcal{I}$. To see how the Gaussian isoperimetric function arises, we Taylor-expand the right-hand side in δ, getting:

$$J(\alpha) + \frac{1}{2J(\alpha)}(J(\alpha)J''(\alpha) + 1)\delta^2 \pm O(\delta^4). \tag{5.6}$$

Thus if take $J = \mathcal{I}$, which satisfies $\mathcal{I}\mathcal{I}'' + 1 = 0$, then the needed inequality (5.5) will at least be satisfied "for small δ, up to an additive $o(\delta^2)$".

Perhaps surprisingly, this is enough to deduce that (5.5) holds exactly, for all δ. This was (in a sense) first established by Barthe and Maurey, who used stochastic calculus and Itô's Formula to prove that (5.5) holds with $J = \mathcal{I}$. Let us present here a sketch of an elementary, discrete version of Barthe–Maurey argument.

We wish to show that Theorem 5.4 holds in the $n = 1$ case; say, for the function $f(\boldsymbol{y}) = \alpha + \beta\boldsymbol{y}$, where $\boldsymbol{y} \sim \{-1, 1\}$. Let's take a random walk on the line, starting from 0, with independent increments $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \dots$ of $\pm\delta$, and stopping when the walk reaches ± 1 (we assume $1/\delta \in \mathbb{Z}^+$). We let $\boldsymbol{y} \in \{-1, 1\}$ be the stopping point of this walk (which is equally likely to be ± 1). Now proving Bobkov's inequality for $f(\boldsymbol{y}) = \alpha + \beta(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3 + \cdots)$ can be reduced to proving Bobkov's inequality just for $f(\boldsymbol{x}_1) = \alpha + \beta\boldsymbol{x}_1$, essentially by the same easy induction used to derive Theorem 5.4 from its $n = 1$ case. This puts us back in the same position as before: we need to show that

$$\mathcal{I}(\alpha) \le \tfrac{1}{2}\sqrt{(\beta\delta)^2 + \mathcal{I}(\alpha+\beta\delta)^2} + \tfrac{1}{2}\sqrt{(\beta\delta)^2 + \mathcal{I}(\alpha-\beta\delta)^2}.$$

However we now have the advantage that the quantity $\beta\delta$ is indeed "small"; we can make it as small as we please. By the Taylor expansion (5.6), the above inequality indeed holds up to an additive $o(\delta^2)$ error. Furthermore, if we simply let this error accumulate in the induction, it costs us almost nothing. It's well known and simple that if $\boldsymbol{T}$ is the number of steps the random walk takes before stopping, then $\mathbf{E}[\boldsymbol{T}] = 1/\delta^2$. Thus we can afford to let an $o(\delta^2)$ error accumulate for $1/\delta^2$ steps, since δ can be made arbitrarily small.

The Barthe–Maurey version of the above argument replaces the random walk with Brownian motion; this is arguably more elegant, but less elementary. An amusing aspect of all this is the following: We first saw in Section 3 that statements about Gaussian geometry can be proven by "simulating" Gaussian random variables by sums of many random ± 1 bits (scaled down); the above argument shows that it can also be effective to simulate a single ± 1 random bit by the sum of many small Gaussians (i.e., with Brownian motion).

We end this section by mentioning that Bobkov's approach to the Gaussian Isoperimetric Inequality inspired Bakry and Ledoux [9, 59] to give a "semigroup proof" of the Gaussian version of Bobkov's inequality (5.3) (á la [8, 58]). Specifically, if one defines

$$F(\rho) = \mathop{\mathbf{E}}_{\gamma_d}[\|(\nabla \mathrm{U}_\rho f,\ \mathcal{I}(\mathrm{U}_\rho f))\|_2],$$

then they showed that F is a nondecreasing function of $\rho \in [0,1]$ just by differentiation (though the computations are a bit cumbersome). This immediately implies (5.3) by taking $\rho = 0, 1$. Mossel and Neeman [65] proved the more general Borell Isoperimetric Inequality

using a very similar semigroup technique, and Ledoux [60] generalized their methodology to include the Hypercontractive Inequality, Brascamp–Lieb inequalities, and some forms of the Slepian inequalities. However, it was by returning to discrete methods — i.e., proving a statement about Boolean functions by induction — that De, Mossel, and Neeman [21] were able to simultaneously establish the Majority Is Stablest Theorem and Borell's theorem.

6. The De–Mossel–Neeman proof of the MIST

Mossel and Neeman actually proved the following functional version of Borell's Isoperimetric Inequality:

Theorem 6.1. *Fix $0 < \rho < 1$ and let $f, g : \mathbb{R}^d \to [0,1]$. Then if $(\boldsymbol{z}, \boldsymbol{z}')$ is a ρ-correlated d-dimensional Gaussian pair,*

$$\mathbf{E}[\Lambda_\rho(f(\boldsymbol{z}), g(\boldsymbol{z}'))] \le \Lambda_\rho(\mathbf{E}[f(\boldsymbol{z})], \mathbf{E}[g(\boldsymbol{z}')]). \tag{6.1}$$

(If $-1 < \rho < 0$ then the inequality is reversed.)

This is equivalent to Borell's inequality in the same way that (5.3) is equivalent to the Gaussian Isoperimetric Inequality (note in particular that $\Lambda_\rho(\alpha, \beta) = \alpha\beta$ when $\alpha, \beta \in \{0, 1\}$). This inequality also has the property that the general-d case follows from the $d = 1$ case by a completely trivial induction, using no special property of Λ_ρ or the Gaussian distribution; it only uses that the d pairs $(\boldsymbol{z}_i, \boldsymbol{z}_i')$ are independent. In particular, *if* (6.1) were to hold for one-bit functions $f, g : \{-1, 1\} \to [0, 1]$ then we could deduce it for general $f, g : \{-1, 1\}^n \to [0, 1]$ by induction, then for Gaussian $f, g : \mathbb{R} \to [0, 1]$ by the Central Limit Theorem, and finally for Gaussian $f, g : \mathbb{R}^d \to [0, 1]$ by induction again. Unfortunately, the inequality (6.1) does *not* hold for $f, g : \{-1, 1\} \to [0, 1]$. It's clear that it can't, because otherwise we would obtain the Majority Is Stablest Theorem with no hypothesis about small influences (which is false). Indeed, the "dictator" functions $f, g : \{-1, 1\} \to [0, 1]$, $f(x) = g(x) = \frac{1}{2} + \frac{1}{2}x$ provide an immediate counterexample; inequality (6.1) becomes the false statement $\frac{1}{4} + \frac{1}{4}\rho \le \frac{1}{2} - \frac{1}{2\pi} \arccos \rho$.

Nevertheless, as noted by De, Mossel, and Neeman [21] we are back in the situation wherein (6.1) "essentially" holds for one-bit functions "with small influences"; i.e., for $f(x) = \alpha + \delta_1 x$, $g(x) = \beta + \delta_2 x$ with δ_1, δ_2 "small". To see this, Taylor-expand the left-hand side of (6.1) around (α, β):

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x}, \boldsymbol{x}') \\ \rho\text{-correlated}}} [\Lambda_\rho(f(\boldsymbol{x}), g(\boldsymbol{x}'))] = \Lambda_\rho(\alpha, \beta) + \mathbf{E}[\delta_1 \boldsymbol{x} \cdot \mathrm{D}_1 \Lambda_\rho(\alpha, \beta)] + \mathbf{E}[\delta_2 \boldsymbol{x}' \cdot \mathrm{D}_2 \Lambda_\rho(\alpha, \beta)]$$
$$+ \mathbf{E}\left[\begin{bmatrix} \delta_1 \boldsymbol{x} & \delta_2 \boldsymbol{x}' \end{bmatrix} \cdot H\Lambda_\rho(\alpha, \beta) \cdot \begin{bmatrix} \delta_1 \boldsymbol{x} \\ \delta_2 \boldsymbol{x}' \end{bmatrix} \right] + \cdots \tag{6.2}$$

(Here $H\Lambda_\rho$ denotes the Hessian of Λ_ρ.) The first term here matches the right-hand side of (6.1). The second and third terms vanish, since $\mathbf{E}[\boldsymbol{x}] = \mathbf{E}[\boldsymbol{x}'] = 0$. Finally, since $\mathbf{E}[\boldsymbol{x}\boldsymbol{x}'] = \rho$ the fourth term is

$$\begin{bmatrix} \delta_1 & \delta_2 \end{bmatrix} \cdot H_\rho \Lambda_\rho(\alpha, \beta) \cdot \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}, \quad \text{where the notation } H_\rho F \text{ means } \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \circ HF. \tag{6.3}$$

One can show by a relatively short calculation that $\det(H_\rho \Lambda_\rho)$ is identically 0 and that the diagonal entries of $H_\rho \Lambda_\rho$ always have opposite sign to ρ. Thus for $0 < \rho < 1$, the matrix $H_\rho \Lambda_\rho$ is everywhere negative semidefinite and hence (6.3) is always nonpositive. (The reverse happens for $0 < \rho < 1$.) Ledoux [60] introduced the terminology *ρ-concavity of F* for the condition $H_\rho F \preccurlyeq 0$.

It follows that (6.1) indeed holds for one-bit Boolean functions f, g, up to the "cubic error term" elided in (6.2). If one now does the induction while keeping these cubic error terms around, the result is the following:

Theorem 6.2 ("De–Mossel–Neeman Theorem"). *Fix $0 < \rho < 1$ and any small $\eta > 0$. Then for $f, g : \{-1,1\}^n \to [\eta, 1-\eta]$,*

$$\begin{aligned} &\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \\ \rho\text{-correlated}}} [\Lambda_\rho(f(\boldsymbol{x}), g(\boldsymbol{y}))] \\ &\le \Lambda_\rho(\mathbf{E}[f(\boldsymbol{x})], \mathbf{E}[g(\boldsymbol{y})]) + O_{\rho,\eta}(1) \cdot \sum_{i=1}^n (\|\mathrm{d}_i f\|_3^3 + \|\mathrm{d}_i g\|_3^3), \end{aligned} \tag{6.4}$$

where $\mathrm{d}_i h$ denotes the ith martingale difference *for h,*

$$(\boldsymbol{x}_1, \dots, \boldsymbol{x}_i) \mapsto \mathbf{E}[h \mid \boldsymbol{x}_1, \dots, \boldsymbol{x}_i] - \mathbf{E}[h \mid \boldsymbol{x}_1, \dots, \boldsymbol{x}_{i-1}].$$

(*For $-1 < \rho < 0$, the inequality* (6.4) *is reversed.*)

With this theorem in hand, Borell's Isoperimetric Inequality for Gaussian functions $f, g : \mathbb{R} \to [\eta, 1-\eta]$ is easily deduced by the standard Central Limit Theorem argument: one only needs to check that the cubic error term is $O(\frac{1}{\sqrt{n}})$, and n may be taken arbitrarily large. Then one immediately deduces the full Borell theorem by taking $\eta \to 0$ and doing another induction on the Gaussian dimension d. On top of this, De, Mossel, and Neeman showed how to deduce Majority Is Stablest from Theorem 6.2, using a small collection of analytical tricks appearing in the original proof. The key trick is to use hypercontractivity to bound $\|\mathrm{d}_i f\|_3^3$ in terms of

$$(\|\mathrm{D}_i f\|_2^2)^{1+\delta} = \mathbf{Inf}_i[f]^{1+\delta}$$

for some small $\delta \approx \frac{\log\log(1/\epsilon)}{\log(1/\epsilon)} > 0$. The fact that we get the nontrivial extra factor $\mathbf{Inf}_i[f]^\delta$, which is at most $\epsilon^\delta \approx \frac{1}{\log(1/\epsilon)}$ by assumption, is the key to finishing the proof.

7. Conclusions: proof complexity

As mentioned, there are two known proofs of the Majority Is Stablest Theorem: the original one, which used the Invariance Principle to reduce the problem to Borell's Isoperimetric Inequality; and, the elegant one due to De, Mossel, and Neeman, which is a completely "discrete proof", as befits a purely discrete problem like Majority Is Stablest. Esthetics is not the only merit of the latter proof, however; as we describe in this section, the fact that the De–Mossel–Neeman proof is simpler and more discrete leads to new technical results concerning the computational complexity of Max-Cut.

Regarding Max-Cut, let's consider the closely related problem of *certifying* that a given graph has no large cut. As we saw in Section 2, for any graph G we can use semidefinite

programming to efficiently compute a value $\beta = \mathrm{SDPOpt}(G)$ such that the maximum cut in G satisfies $\mathrm{Opt}(G) \le \beta$. We think of this algorithm as producing a *proof* of the statement "$\mathrm{Opt}(G) \le \beta$". Furthermore, the (analysis of the) Goemans–Williamson algorithm implies that the bound found by this algorithm is fairly good; whenever G truly satisfies $\mathrm{Opt}(G) \le \frac{\theta}{\pi}$ (for $\theta \in [\theta_{\mathrm{GW}}, \pi]$), we will efficiently obtain a proof of "$\mathrm{Opt}(G) \le \frac{1}{2} - \frac{1}{2}\cos\theta$". For example, if $\mathrm{Opt}(G) \le \frac{3}{4}$ then there is an efficiently-obtainable "SDP proof" of the statement "$\mathrm{Opt}(G) \le \frac{1}{2} + \frac{1}{2\sqrt{2}} \approx .854$".

Assuming the Unique Games Conjecture (and P $\neq$ NP), the works [51, 66] imply that there is no efficient algorithm that can in general find better proofs; e.g., that can certify "$\mathrm{Opt}(G) \le .853$" whenever $\mathrm{Opt}(G) \le \frac{3}{4}$. In fact, under the additional standard assumption of coNP $\neq$ NP, the implication is simply that no short proofs *exist*; i.e., there are infinite families of graphs $G = (V, E)$ with $\mathrm{Opt}(G) \le \frac{3}{4}$ but no $\mathrm{poly}(|V|)$-length proof of the statement "$\mathrm{Opt}(G) \le .853$" (say, in some textbook formalization of mathematical reasoning). In other words:

Unique Games & P $\neq$ NP prediction about Max-Cut: Let $\theta \in [\theta_{\mathrm{GW}}, \pi]$ and $\delta > 0$. There is no polynomial-time algorithm that, given a Max-Cut instance G with $\mathrm{Opt}(G) \le \frac{\theta}{\pi}$, outputs a proof of "$\mathrm{Opt}(G) \le \frac{1}{2} - \frac{1}{2}\cos\theta - \delta$".

Unique Games & coNP $\neq$ NP prediction about Max-Cut: In fact, there are infinitely many graphs G with $\mathrm{Opt}(G) \le \frac{\theta}{\pi}$, yet for which no polynomial-length proof of "$\mathrm{Opt}(G) \le \frac{1}{2} - \frac{1}{2}\cos\theta - \delta$" exists.

As mentioned, the Unique Games Conjecture is quite contentious, so it's important to seek additional evidence concerning the above predictions. For example, to support the first prediction one should at a minimum show that the semidefinite program (SDP) fails to provide such proofs. That is, one should find graphs G with $\mathrm{Opt}(G) \le \frac{\theta}{\pi}$ yet $\mathrm{SDPOpt}(G) \ge \frac{1}{2} - \frac{1}{2}\cos\theta$. Such graphs are called *SDP integrality gap instances*, as they exhibit a large gap between their true optimal Max-Cut and the upper-bound certified by the SDP. Borell's Isoperimetric Inequality precisely provides such graphs, at least if "weighted continuous graphs" are allowed: One takes the "graph" G whose vertex set is $\mathbb{R}^d$ and whose "edge measure" is given by choosing a $(\cos\theta)$-correlated pair of Gaussians. The fact that $\mathrm{Opt}(G) \le \frac{\theta}{\pi}$ is immediate from Borell's Theorem 3.4; further, it's not hard to show (using the idea of Remark 3.2) that choosing $\vec{U}(v) = v/\sqrt{d}$ in (SDP) establishes $\mathrm{SDPOpt}(G) \ge \frac{1}{2} - \frac{1}{2}\cos\theta - o_d(1)$. These facts were essentially established originally by Feige and Schechtman [30], who also showed how to discretize the construction so as to provide finite integrality gap graphs.

(Incidentally, we may now explain that the Raghavendra Theory mentioned at the end of Section 2 significantly generalizes the work of Khot et al. [51] by showing how to transform an SDP integrality gap instance for *any* CSP into a matching computational hardness-of-approximation result, assuming the Unique Games Conjecture.)

Although the semidefinite program (SDP) fails to certify $\mathrm{Opt}(G) \le \frac{\theta}{\pi}$ for the "correlated Gaussian graphs" described above, a great deal of recent research has gone into developing stronger "proof systems" for reasoning about Max-Cut and other CSPs. (See, e.g., [33] for a survey.) Actually, until recently this research was viewed not in terms of proof complexity but in terms of analyzing "tighter" SDP relaxations that can still be solved efficiently. For

example, one can still solve the optimization problem (SDP) in polynomial time with the following "triangle inequality" constraint added in:

$$\langle U(v), U(w)\rangle + \langle U(w), U(x)\rangle - \langle U(v), U(x)\rangle \le 1 \quad \forall v, w, x \in V.$$

Note that with this additional constraint we still have $\mathrm{Opt}(G) \le \mathrm{SDPOpt}(G)$ for all G, because the constraint is satisfied by any genuine bipartition $U : V \to \{-1, 1\}$. As noted by Feige and Schechtman [30], adding this constraint gives a certification better than "$\mathrm{Opt}(G) \le \frac{1}{2} - \frac{1}{2}\cos\theta$" for the Gaussian correlation graphs, though it's not clear by how much.

Although this stronger "SDP + triangle inequality" proof system does better on Gaussian correlation graphs, a breakthrough work of Khot and Vishnoi [54] showed that it still suffers from the same integrality gap for a different infinite family of graphs. In other words, even when the SDP includes the triangle inequalities, these *Khot–Vishnoi graphs* $G = (V, E)$ have $\mathrm{SDPOpt}(G) \ge \frac{1}{2} - \frac{1}{2}\cos\theta$ yet $\mathrm{Opt}(G) \le \frac{\theta}{\pi} + o_{|V|}(1)$. The second fact, the upper bound on the true Max-Cut value, relies directly on the Majority Is Stablest Theorem. Subsequent works [53, 75] significantly generalized this result by showing that even much tighter "SDP hierarchies" still fail to certify anything better than "$\mathrm{Opt}(G) \le \frac{1}{2} - \frac{1}{2}\cos\theta$" for the Khot–Vishnoi graphs G. This could be considered additional evidence in favor of the Unique Games & P $\neq$ NP Prediction concerning Max-Cut.

A recent work by Barak et al. [11] cast some doubt on this prediction, however. Their work showed that the especially strong "Lasserre/Parrilo SDP hierarchy" [57, 72, 79] succeeds in finding some good CSP bounds which weaker SDP hierarchies are unable to obtain. Specifically, they showed it provides good upper bounds on the optimal value of the Khot–Vishnoi "Unique Games instances" (which are, in some sense, subcomponents of the Khot–Vishnoi Max-Cut graphs). Subsequent work of O'Donnell and Zhou [71] further emphasized the equivalence of the Lasserre/Parrilo SDP hierarchy and the *Sum-of-Squares (SOS) proof system*, invented by Grigoriev and Vorobjov [36]. In the context of the Max-Cut CSP, this proof system (inspired by Hilbert's 17th Problem [42] and the *Positivstellensatz* of Krivine [56] and Stengle [80]) seeks to establish the statement "$\mathrm{Opt}(G) \le \beta$" for a graph $G = (V, E)$ by expressing

$$\beta - \left(\operatorname*{avg}_{(v,w)\in E} \tfrac{1}{2} - \tfrac{1}{2} X_v X_w \right) = \sum_{i=1}^{s} P_i^2 \quad \text{within the ring } \mathbb{R}[(X_v)_{v\in V}]/(X_v^2 - 1)_{v\in V}, \tag{7.1}$$

for some formal polynomials $P_1, \ldots, P_s$ of degree at most some constant C. Somewhat remarkably, there is an efficient ($|V|^{O(C)}$-time) algorithm for finding such P_i's whenever they exist.

As mentioned, for the Khot–Vishnoi Max-Cut graphs G, the fact that $\mathrm{Opt}(G) \le \frac{\theta}{\pi} + o(1)$ follows directly from the Majority Is Stablest Theorem. To show that the SOS proof system can also certify this fact (thereby casting some doubt on the Unique Games & P $\neq$ NP Prediction about Max-Cut), one needs to show that not only is the Majority Is Stablest Theorem true, but that it can be proved within the extremely constrained SOS proof system, á la (7.1). The original proof of the Majority Is Stablest Theorem was quite complicated, using the Invariance Principle from [67] to reduce Borell's Isoperimetric Inequality, and then relying on the known geometric proofs [12, 17] of the latter. The prospect for converting this proof into an SOS format seemed quite daunting (although a partial result was established in [71], showing that the SOS proof system can establish "$\mathrm{Opt}(G) \le \frac{1}{2} - \frac{\cos\theta}{\pi} - (\frac{1}{2} -$

$\frac{1}{\pi})\cos^3\theta$"). However, the simplicity and discrete nature of the new De–Mossel–Neeman proof of the Majority Is Stablest Theorem allowed them to show that the SOS proof system *can* establish the truth about the Khot–Vishnoi graphs, $\mathrm{Opt}(G) \leq \frac{1}{2} - \frac{1}{2}\cos\theta + o(1)$.

It is to be hoped that this result can be extended to the entire Raghavendra Theory, thereby showing that the SOS proof system can certify the optimal value of the analogue of the Khot–Vishnoi instances for *all* CSPs. However as the Raghavendra Theory still relies on the Invariance Principle, whether or not this is possible is unclear.

Finally, in light of the De–Mossel-Neeman result, the following interesting question is open: Are there (infinite families of) instances of the Max-Cut problem G such that $\mathrm{Opt}(G) \leq \frac{\theta}{\pi}$, yet such that any mathematical proof of this statement is so complicated that the SOS proof system cannot establish anything better than "$\mathrm{Opt}(G) \leq \frac{1}{2} - \frac{1}{2}\cos\theta$"? If such graphs were found, this might tilt the weight of evidence back in favor of the Unique Games & P $\neq$ NP Prediction. Of course, if human mathematicians explicitly construct the proof of $\mathrm{Opt}(G) \leq \frac{\theta}{\pi}$, presumably it will have polynomial length, and therefore not provide any evidence in favor of the Unique Games & coNP $\neq$ NP Prediction. To provide evidence for this stronger prediction, one presumably needs to give a *probabilistic* construction of graphs G such that both of the following happen with high probability: (i) $\mathrm{Opt}(G) \leq \frac{\theta}{\pi}$; and, (ii) there is no polynomial-length proof even of "$\mathrm{Opt}(G) \leq \frac{1}{2} - \frac{1}{2}\cos\theta$".

Acknowledgements. The author is supported by NSF grants CCF-1319743 and CCF-1116594.

References

[1] L. Ambrosio and A. Figalli, *Surface measures and convergence of the Ornstein–Uhlenbeck semigroup in Wiener spaces*, Annales de la faculté des sciences de Toulouse Mathématiques (série 6), **20**(2) (2011), 407–438.

[2] L. Ambrosio, A. Figalli, and E. Runa, *On sets of finite perimeter in Wiener spaces: reduced boundary and convergence to halfspaces*, Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Serie IX. Matematica e Applicazioni, **24**(1) (2013), 111–122.

[3] L. Ambrosio, M. Miranda Jr., S. Maniglia, and D. Pallara, *BV functions in abstract Wiener spaces*, Journal of Functional Analysis, **258**(3) (2010), 785–813.

[4] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, *Proof verification and the hardness of approximation problems*, Journal of the ACM, **45**(3) (1998), 501–555.

[5] S. Arora and S. Safra, *Probabilistic checking of proofs: A new characterization of NP*, Journal of the ACM, **45**(1) (1998), 70–122.

[6] K. Arrow, *A difficulty in the concept of social welfare*, The Journal of Political Economy, **58**(4) (1950), 328–346.

[7] A. Baernstein and B. Taylor, *Spherical rearrangements, subharmonic functions, and $*$-functions in n-space*, Duke Mathematical Journal, **43**(2) (1976), 245–268.

[8] D. Bakry and M. Émery, *Diffusions hypercontractives*, In Séminaire de Probabilités, XIX, volume 1123 of Lecture Notes in Mathematics, Springer, Berlin, 1985, pp. 177–206.

[9] D. Bakry and M. Ledoux, *Lévy–Gromov's isoperimetric inequality for an infinite dimensional diffusion generator*, Inventiones mathematicae, **123**(1) (1996), 259–281.

[10] J. Banzhaf, *Weighted voting doesn't work: A mathematical analysis*, Rutgers Law Review, **19** (1965), 317–343.

[11] B. Barak, F. Brandão, A. Harrow, J. Kelner, D. Steurer, and Y. Zhou, *Hypercontractivity, sum-of-squares proofs, and their applications*, In Proceedings of the 44th Annual ACM Symposium on Theory of Computing, 2012, pp. 307–326.

[12] W. Beckner, *Sobolev inequalities, the Poisson semigroup, and analysis on the sphere* S^n, Proceedings of the National Academy of Sciences, **89**(11) (1992), 4816–4819.

[13] M. Bellare, O. Goldreich, and M. Sudan, *Free bits, PCPs, and non-approximability – towards tight results*, SIAM Journal of Computing, **27**(3) (1998), 804–915.

[14] E. Blais and R. O'Donnell, *Lower bounds for testing function isomorphism*, In Proceedings of the 25th Annual IEEE Conference on Computational Complexity, 2010, pp. 235–246.

[15] A. Bonami, *Étude des coefficients Fourier des fonctions de* $L^p(G)$, Annales de l'Institut Fourier, **20**(2) (1970), 335–402.

[16] C. Borell, *The Brunn–Minkowski inequality in Gauss space*, Inventiones Mathematicae, **30**,(2) (1975), 207–216.

[17] ———, *Geometric bounds on the Ornstein–Uhlenbeck velocity process*, Probability Theory and Related Fields, **70**(1) (1985), 1–13.

[18] S. Brams, W. Gehrlein, and F. Roberts, editors, *The Mathematics of Preference, Choice and Order*, Springer, 2009.

[19] E. Carlen and M. Loss, *Extremals of functionals with competing symmetries*, Journal of Functional Analysis, **88**(2) (1990), 437–456.

[20] A. De, E. Mossel, and J. Neeman, *Majority is Stablest : discrete and SoS*, Technical Report arXiv:1211.1001, 2012.

[21] A. De, E. Mossel, and J. Neeman, *Majority is Stablest : Discrete and SoS*, In Proceedings of the 45th Annual ACM Symposium on Theory of Computing, 2013.

[22] N. de Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris, de l'imprimerie royale, 1785.

[23] C. Delorme and S. Poljak, *Laplacian eigenvalues and the maximum cut problem*, Mathematical Programming, **62**(1–3) (1993), 557–574.

[24] I. Dinur, E. Friedgut, and O. Regev, *Independent sets in graph powers are almost contained in juntas*, Geometric and Functional Analysis, **18**(1) (2008), 77–97.

[25] A. Ehrhard, *Symétrisation dans l'espace de gauss*, Mathematica Scandinavica, **53** (1983), 281–301.

[26] A. Ehrhard, *Inégalités isopérimétriques et intégrales de Dirichlet gaussiennes*, Annales Scientifiques de l'École Normale Supérieure. Quatrième Série, **17**(2) (1984), 317–332.

[27] R. Eldan, *A two-sided estimate for the Gaussian noise stability deficit*, Technical Report arXiv:1307.2781, 2013.

[28] U. Feige, *Randomized rounding of semidefinite programs – variations on the Max-Cut example*, volume 1761 of Lecture Notes in Computer Science, Springer, 1999, pp. 189–196.

[29] U. Feige, S. Goldwasser, L. Lovász, S. Safra, and M. Szegedy, *Interactive proofs and the hardness of approximating cliques*, Journal of the ACM, **43**(2) (1996), 268–292.

[30] U. Feige and G. Schechtman, *On the optimality of the random hyperplane rounding technique for Max-Cut*, Randoom Structures and Algorithms, **20**(3) (2002), 403–440.

[31] E. Friedgut, G. Kalai, and A. Naor, *Boolean functions whose Fourier transform is concentrated on the first two levels and neutral social choice*, Advances in Applied Mathematics, **29**(3) (2002), 427–437.

[32] M. Garman and M. Kamien, *The paradox of voting: probability calculations*, Behavioral Science, **13**(4) (1968), 306–316.

[33] K. Georgiou, *Integrality gaps for strong linear programming and semidefinite programming relaxations*, PhD thesis, University of Toronto, 2010.

[34] M. Goemans and D. Williamson, *A 0.878 approximation algorithm for MAX-2SAT and MAX-CUT*, In Proceedings of the 26th Annual ACM Symposium on Theory of Computing, 1994, pp. 422–431.

[35] M. Goemans and D. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, Journal of the ACM, **42** (1995), 1115–1145.

[36] D. Grigoriev and N. Vorobjov, *Complexity of Null- and Positivstellensatz proofs*, Annals of Pure and Applied Logic, **113**(1) (2001), 153–160.

[37] L. Gross, *Logarithmic Sobolev inequalities*, American Journal of Mathematics, **97**(4) (1975), 1061–1083.

[38] G.-T. Guilbaud, *Les théories de l'intérêt général et le problème logique de l'agrégation*, Economie appliquée, V(4) (1952), 501–551.

[39] L. Harper, *Optimal assignments of numbers to vertices*, Journal of the Society for Industrial and Applied Mathematics, **12**(1) (1964), 131–135.

[40] J. Håstad, *Some optimal inapproximability results*, In Proceedings of the 29th Annual ACM Symposium on Theory of Computing, 1997, pp. 1–10.

[41] _____, *Some optimal inapproximability results*, Journal of the ACM, **48**(4) (2001), 798–859.

[42] D. Hilbert, *Mathematical problems*, Bulletin of the American Mathematical Society, **8**(10) (1902), 437–479.

[43] M. Hino, *Sets of finite perimeter and the Hausdorff-Gauss measure on the Wiener space*, Journal of Functional Analysis, **258**(5) (2010), 1656–1681.

[44] S. Janson, *Gaussian Hilbert Spaces*, Cambridge University Press, 1997.

[45] D. Johnson, *The NP-Completeness column: the many limits on approximation*, ACM Transactions on Algorithms, **2**(3) (2006), 473–489.

[46] G. Kalai, *A Fourier-theoretic perspective on the Condorcet paradox and Arrow's theorem*, Advances in Applied Mathematics, **29**(3) (2002), 412–426.

[47] D. Kane, *The correct exponent for the Gotsman–Linial conjecture*, Technical Report arXiv:1210.1283, 2012.

[48] H. Karloff, *How good is the Goemans–Williamson MAX CUT algorithm?*, SIAM Journal of Computing, **29**(1) (1999), 336–350.

[49] R. Karp, *Reducibility among combinatorial problems*, In Complexity of Computer Computations, Plenum Press, 1972, pp. 85–103.

[50] S. Khot, *On the power of unique 2-prover 1-round games*, In Proceedings of the 34th Annual ACM Symposium on Theory of Computing, 2002, pp. 767–775.

[51] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell, *Optimal inapproximability results for MAX-CUT and other 2-variable CSPs?*, In Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science, 2004, pp. 146–154.

[52] _____, *Optimal inapproximability results for Max-Cut and other* 2*-variable CSPs?*, SIAM Journal on Computing, **37**(1) (2007), 319–357.

[53] S. Khot and R. Saket, *SDP integrality gaps with local ℓ_1-embeddability*, In Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science, 2009, pp. 565–574.

[54] S. Khot and N. Vishnoi, *The Unique Games Conjecture, integrality gap for cut problems and embeddability of negative type metrics into ℓ_1*, In Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science, 2005, pp. 53–62.

[55] G. Kindler and R. O'Donnell, *Gaussian noise sensitivity and Fourier tails*, In Proceedings of the 26th Annual IEEE Conference on Computational Complexity, 2012, pp. 137–147.

[56] J.-L. Krivine, *Anneaux préordonnés*, Journal d'Analyse Mathématique, **12**(1) (1964), 307–326.

[57] J. Lasserre, *Optimisation globale et théorie des moments*, Comptes Rendus de l'Académie des Sciences, **331**(11) (2000), 929–934.

[58] M. Ledoux, *Semigroup proofs of the isoperimetric inequality in Euclidean and Gauss space*, Bulletin des Sciences Mathématiques, **118**(6) (1994), 485–510.

[59] _____, *A short proof of the Gaussian isoperimetric inequality*, In High dimensional probability (Oberwolfach, 1996), volume 43 of Progress in Probability, Birkhäuser, Basel, 1998, pp. 229–232.

[60] _____, *Remarks on noise sensitivity, Brascamp–Lieb and Slepian inequalities*, http://perso.math.univ-toulouse.fr/ledoux/files/2013/11/noise.pdf, 2013.

[61] P. Lévy, *Leçons d'Analyse Fonctionnelle*, Gauthier-Villars, 1922.

[62] R. Meka and D. Zuckerman, *Pseudorandom generators for polynomial threshold functions*, In Proceedings of the 42nd Annual ACM Symposium on Theory of Computing, 2010, pp. 427–436.

[63] M. Miranda Jr., M. Novaga, and D. Pallara, *An introduction to BV functions in Wiener spaces*, Technical Report arXiv:1212.5926, 2012.

[64] E. Mossel, *Gaussian bounds for noise correlation of functions*, Geometric and Functional Analysis, **19**(6) (2010), 1713–1756.

[65] E. Mossel and J. Neeman, *Robust optimality of Gaussian noise stability*, Technical Report arXiv:1210.4126, 2012.

[66] E. Mossel, R. O'Donnell, and K. Oleszkiewicz, *Noise stability of functions with low influences: invariance and optimality*, In Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science, 2005, pp. 21–30.

[67] _____, *Noise stability of functions with low influences: invariance and optimality*, Annals of Mathematics, **171**(1) (2010), 295–341.

[68] E. Mossel, R. O'Donnell, O. Regev, J. Steif, and B. Sudakov, *Non-interactive correlation distillation, inhomogeneous Markov chains, and the reverse Bonami–Beckner inequality*, Israel Journal of Mathematics, **154** (2006), 299–336.

[69] R. O'Donnell, *Analysis of Boolean Functions*, Cambridge University Press, 2014.

[70] R. O'Donnell and Y. Wu, *An optimal SDP algorithm for Max-Cut, and equally optimal Long Code tests*, In Proceedings of the 40th Annual ACM Symposium on Theory of Computing, 2008, pp. 335–344.

[71] R. O'Donnell and Y. Zhou, *Approximability and proof complexity*, In Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms, 2013, pp. 1537–1556.

[72] P. Parrilo, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, PhD thesis, California Institute of Technology, 2000.

[73] L. Penrose, *The elementary statistics of majority voting*, Journal of the Royal Statistical Society, **109**(1) (1946), 53–57.

[74] P. Raghavendra, *Optimal algorithms and inapproximability results for every CSP?*, In Proceedings of the 40th Annual ACM Symposium on Theory of Computing, 2008, pp. 245–254.

[75] P. Raghavendra and D. Steurer, *Integrality gaps for strong SDP relaxations of Unique Games*, In Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science, 2009, pp. 575–585.

[76] V. Rotar', *Limit theorems for multilinear forms and quasipolynomial functions*, Teoriya Veroyatnostei i ee Primeneniya, **20**(3) (1975), 527–546.

[77] E. Schmidt, *Die Brunn-Minkowskische Ungleichung und ihr Spiegelbild sowie die isoperimetrische Eigenschaft der Kugel in der euklidischen und nichteuklidischen Geometrie. I*, Mathematische Nachrichten, **1** (1948), 81–157.

[78] W. Sheppard, *On the application of the theory of error to cases of normal distribution and normal correlation*, Philosophical Transactions of the Royal Society of London, Series A, **192** (1899), 101–167, 531.

[79] N. Shor, *Class of global minimum bounds of polynomial functions*, Cybernetics, **23**(6) (1987), 731–734.

[80] G. Stengle, *A Nullstellensatz and a Positivstellensatz in semialgebraic geometry*, Mathematische Annalen, **207**(2) (1973), 87–97.

[81] V. Sudakov and B. Tsirel'son, *Extremal properties of half-spaces for spherically invariant measures*, Journal of Soviet Mathematics, **9**(1) (1978), 9–18, Originally published in Russian in 1974.

[82] L. Trevisan, G. Sorkin, M. Sudan, and D. Williamson, *Gadgets, approximation, and linear programming*, SIAM Journal on Computing, **29**(6) (2000), 2074–2097.

Computer Science Department, Carnegie Mellon University, Pittsburgh, USA
E-mail: odonnell@cs.cmu.edu

Algorithms for circuits and circuits for algorithms: Connecting the tractable and intractable

Ryan Williams

Abstract. The title of this paper highlights an emerging duality between two basic topics in algorithms and complexity theory. *Algorithms for circuits* refers to the design of algorithms which can analyze finite logical circuits or Boolean functions as input, checking a simple property about the complexity of the underlying function. For instance, an algorithm determining if a given logical circuit C has an input that makes C output *true* would solve the NP-complete Circuit-SAT problem. Such an algorithm is unlikely to run in polynomial time, but could possibly be more efficient than exhaustively trying all possible inputs to the circuit. *Circuits for algorithms* refers to the modeling of "complex" uniform algorithms with "simple" Boolean circuit families, or proving that such modeling is impossible. For example, can every exponential-time algorithm be simulated using Boolean circuit families of only polynomial size? It is widely conjectured that the answer is *no*, but the present mathematical tools available are still too crude to resolve this kind of separation problem. This paper surveys these two generic subjects and the connections that have been developed between them, focusing on connections between non-trivial circuit-analysis algorithms and proofs of circuit complexity lower bounds.

Mathematics Subject Classification (2010). Primary 68Q17; Secondary 68Q25.

Keywords. circuit complexity, algorithm analysis, satisfiability, lower bounds, derandomization, learning, exact algorithms, parameterized algorithms.

1. Introduction

Budding theoretical computer scientists are generally taught several dictums at an early age. One such dictum is that the algorithm designers and the complexity theorists (whoever they may be) are charged with opposing tasks. The algorithm designer discovers interesting methods for solving certain problems; along the way, she may also propose new notions of what is interesting, to better understand the scope and power of algorithms. The complexity theorist is supposed to prove *lower bounds*, showing that sufficiently interesting methods for solving certain problems do not exist. Barring that, he develops a structural framework that explains the consequences of such impossibility results, as well as consequences of possessing such interesting methods.

Another dictum is that algorithm design and analysis is, on the whole, an easier venture than proving lower bounds. In algorithm design, one only has to find a single efficient algorithm that will solve the problem at hand, but a lower bound must reason about *all* possible efficient algorithms, including bizarrely behaving ones, and argue that none solve the problem at hand. This dictum is also reflected in the literature: every year, many interest-

Proceedings of the International Congress of Mathematicians, Seoul, 2014

ing algorithms are discovered, analyzed, and published, compared to the tiny number of lower bounds proved.[1] Furthermore, there are rigorously mathematical reasons for believing that lower bounds are hard to prove. The most compelling of these are the three "barriers" of Relativization [9], Natural Proofs [64], and Algebrization [1]. These "no-go" theorems demonstrate that the known lower bound proof methods are simply too coarse to prove even weak lower bounds, much weaker than $\mathsf{P} \neq \mathsf{NP}$. Subsequently, complexity theory has been clouded with great pessimism about resolving some of its central open problems.

While the problems of algorithm design and proving lower bounds may arise from looking at opposing tasks, the two tasks do have deep *similarities* when viewed in the appropriate way.[2] This survey will concentrate on some of the most counterintuitive similarities: from the design of certain algorithms (the supposedly "easier" task), one can derive new lower bounds (the supposedly "harder" task). That is, there are senses in which algorithm design is *at least as hard* as proving lower bounds, contrary to dictums. These connections present an excellent mathematical "arbitrage" opportunity for complexity theorists: to potentially prove hard lower bounds via supposedly easier algorithm design. (Moreover, there is money to be made: this approach *has* recently led to new lower bounds.)

Several connections take the following form:

> The *existence* of an "efficient" algorithm T that can analyze *all* structured circuits C implies the *existence* of an "efficient" function f that is not computable by *all* structured circuit families.

Therefore, while algorithms and lower bounds are opposites by definition, there are situations where algorithm design for a problem X can be translated into "lower bound design" for another problem Y. The key is that there are two computational models under consideration here: the *algorithm model* or the usual "Turing" style model of algorithms, and the *circuit model* or the non-uniform circuit family model, which we shall define shortly. Careful design of algorithms for analyzing instances of the circuit model are used to construct functions computable (in one sense) in the algorithm model that are uncomputable (in another sense) in the circuit model. There is a kind of duality lurking beneath which is not well-understood.

The focus of this article is on two generic topics in algorithms and complexity, and connections between them:

- *Circuits for Algorithms* refers to the modeling of powerful uniform algorithms with non-uniform circuit families, or proving that such modeling is impossible. For instance, the celebrated EXP versus $\mathsf{P/poly}$ question asks if exponential-time algorithms can be simulated using non-uniform circuit families of polynomial size. Complexity theorists believe that the answer is *no*, but they presently have no idea how to prove such a circuit lower bound.
- *Algorithms for Circuits* refers to the design of algorithms which can analyze finite logical circuits or Boolean functions as input, checking some property about the complexity of the underlying function. To illustrate, the problem Circuit-SAT asks if a given logical circuit has an input that forces the circuit to output *true*. Circuit-SAT

[1] Of course, there can be other reasons for this disparity, such as funding.

[2] Similarities are already present in the proof(s) that the Halting Problem is undecidable: such results rely on the construction of a *universal Turing machine* that can run arbitrary Turing machine code given as input. This is a textbook application of how an algorithm can be used to prove an impossibility theorem.

is NP-complete and believed to be intractable; nevertheless, even "mildly intractable" algorithms for this problem would be useful in both theory and practice. It is an outstanding open question whether one can asymptotically improve over the "brute force" algorithm for Circuit-SAT which simply evaluates the circuit on all possible inputs. Recent surprising developments have shown that even tiny improvements over exhaustive search would significantly impact Circuits for Algorithms—in fact, new circuit lower bounds have been deduced from such algorithms.

The rest of the paper is organized as follows. The next section provides a bit of relevant background. Section 3 surveys *circuits for algorithms*, and Section 4 surveys *algorithms for circuits*. Section 5 discusses known connections between the two, and prospects for future progress. Section 6 briefly concludes.

2. Preliminaries

Recall $\{0,1\}^n$ is the set of all n-bit binary strings, and $\{0,1\}^\star = \bigcup_{n\in\mathbb{N}}\{0,1\}^n$.

A quick recollection of machine-based complexity Any reasonable algorithmic model with a coherent method for counting steps (such as Turing machines and their transition functions) will suffice for our discussion. For an algorithm A, we let $A(x)$ denote the output of A on the input x. A *language* L is a subset of $\{0,1\}^\star$; in the following, the variable L always denotes a language. We typically think of L as an indicator function from $\{0,1\}^\star$ to $\{0,1\}$, in the natural way.

Let $t : \mathbb{N} \to \mathbb{N}$. An algorithm A *runs in time* $t(n)$ if, on all $x \in \{0,1\}^n$, $A(x)$ halts within $t(|x|)$ steps. *Decidability of L in time* $t(n)$ means that there is an algorithm A running in time $t(n)$ such that $A(x) = L(x)$ for all x.

L is *verifiable in time* $t(n)$ if there exists an algorithm A such that, on all $x \in \{0,1\}^n$, $x \in L$ if and only if there is a $y_x \in \{0,1\}^{t(|x|)}$ such that $A(x, y_x)$ runs in time $O(t(|x|))$ and $A(x, y_x) = 1$. Intuitively, the string y_x serves as a *proof* that $x \in L$, and this proof can be verified in time $O(t(|x|))$.

An algorithm A *runs in space* $t(n)$ if, on all $x \in \{0,1\}^n$, the total workspace used by $A(x)$ is at most $t(|x|)$ cells (or registers, or bits, depending on the model). Decidability of a language in space $t(n)$ is defined in the obvious way.

Some complexity classes relevant to our discussion are:

- P: the class of languages decidable in $O(p(n))$ time for some $p \in \mathbb{Z}[x]$.
- NP: languages verifiable in $O(p(n))$ steps for some $p \in \mathbb{Z}[x]$.
- PSPACE: languages decidable in space $O(p(n))$ for some $p \in \mathbb{Z}[x]$.
- EXP: languages decidable in $O(2^{p(n)})$ time for some $p \in \mathbb{Z}[x]$.
- NEXP: languages verifiable in $O(2^{p(n)})$ time for some $p \in \mathbb{Z}[x]$.
- EXPSPACE: languages decidable in space $O(2^{p(n)})$ for some $p \in \mathbb{Z}[x]$.

Let C be one of the above classes. An *algorithm A with oracle access to C* has a powerful extra instruction: there is a language $L \in C$ such that A can call $L(y)$ in one time step, on any input y of its choosing. (Intuitively, A can efficiently "consult the oracle" in class C for

answers.) This is an interesting notion when C is a hard complexity class, say in NP or in PSPACE, and L is chosen to be a hard language in C.

Circuit complexity A function $f : \{0,1\}^n \to \{0,1\}$ is called *Boolean*. We let $x_1, \ldots, x_n$ denote the n variables to a Boolean function f. Circuit complexity is chiefly concerned with the difficulty of building up Boolean functions out of "simpler" functions, such as those of the form $g : \{0,1\}^2 \to \{0,1\}$. Examples of interesting Boolean functions include:

- $\text{OR}_k(x_1, \ldots, x_k)$, $\text{AND}_k(x_1, \ldots, x_k)$, with their usual logical meanings,
- $\text{MODm}_k(x_1, \ldots, x_k)$ for a fixed integer $m > 1$, which outputs 1 if and only if $\sum_i x_i$ is divisible by m.
- $\text{MAJ}_k(x_1, \ldots, x_k) = 1$ if and only if $\sum_i x_i \geq \lceil k/2 \rceil$.

A *basis set* $\mathcal{B}$ is a set of Boolean functions. Two popular choices for $\mathcal{B}$ are B_2, the set of all functions $g : \{0,1\}^2 \to \{0,1\}$, and U_2, the set B_2 without MOD2 and the negation of MOD2. A *Boolean circuit* of *size* s with n inputs $x_1, \ldots, x_n$ over basis $\mathcal{B}$ is a sequence of $n + s$ functions $C = (f_1, \ldots, f_{n+s})$, with $f_i : \{0,1\}^n \to \{0,1\}$ for all i, such that:

- for all $i = 1, \ldots, n$, $f_i(x_1, \ldots, x_n) = x_i$,
- for all $j = n+1, \ldots, n+s$, there is a function $g : \{0,1\}^k \to \{0,1\}$ from $\mathcal{B}$ and indices $i_1, \ldots, i_k < j$ such that

$$f_j(x_1, \ldots, x_n) = g(f_{i_1}(x_1, \ldots, x_n), \ldots, f_{i_k}(x_1, \ldots, x_n)).$$

The f_i are the *gates* of the circuit; $f_1, \ldots, f_n$ are the *input gates*, $f_{n+1}, \ldots, f_{n+s-1}$ are the *internal gates*, and f_{n+s} is the *output gate*. The circuit C can naturally be thought of as a function as well: on an input string $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, $C(x)$ denotes $f_{n+s}(x)$.

Thinking of the connections between the gates as a directed acyclic graph in the natural way, with the input gates as n source nodes $1, \ldots, n$, and the jth gate with indices $i_1, \ldots, i_k < j$ as a node j with incoming arcs from nodes $i_1, \ldots, i_k$, the *depth* of C is the longest path from an input gate to the output gate. As a convention, gates with fan-in 1 are not counted in the depth measure. That is, gates of the form $g(x) = x_i$ or $g(x) = \neg x_i$ are not counted towards the length of a path from input to output.

Given a basis set $\mathcal{B}$ and $f : \{0,1\}^n \to \{0,1\}$, what is the minimal size s of a Boolean circuit over $\mathcal{B}$ with output gate $f_{n+s} = f$? This quantity is the *$\mathcal{B}$-circuit complexity of f*, and is denoted by $C_{\mathcal{B}}(f)$. The minimal depth of a circuit computing f is also of interest for parallel computing, and is denoted by $D_{\mathcal{B}}(f)$.

3. Circuits for algorithms

The circuit model is excellent for understanding the difficulty and efficiency of computing *finite* functions. For every $f : \{0,1\}^n \to \{0,1\}$ and basis set, the circuit complexity of f is a fixed integer which could be high or low, relative to n.

Boolean circuits should be contrasted with the typical *uniform algorithm* models used in computability and complexity theory, based on finite objects such as Turing machines. In

that setting, one is presented with functions (languages) defined over infinitely many strings, i.e., of the form

$$L : \{0,1\}^\star \to \{0,1\}, \tag{3.1}$$

and a primary goal is to find a fixed program or machine M such that, for every input $x \in \{0,1\}^\star$, running M on input x always produces the output $L(x)$ in some finite (or efficient) number of steps. This sort of computational model can trivially compute all finite functions (outputting 1 on only finitely many inputs) in *constant* time, by hard-coding the answers to each of the finitely many inputs in the program's code.

There is a logical way to extend the Boolean circuit model to also compute functions of type (3.1): we simply provide infinitely many circuits.

Definition 3.1. Let $s : \mathbb{N} \to \mathbb{N}$, $d : \mathbb{N} \to \mathbb{N}$, and $L : \{0,1\}^\star \to \{0,1\}$. *$L$ has size-$s(n)$ depth-$d(n)$ circuits* if there is an infinite family $\{C_n \mid n \in \mathbb{N}\}$ of Boolean circuits over B_2 such that, for every n, C_n has n inputs, size at most $s(n)$, depth at most $d(n)$, and for all $x \in \{0,1\}^n$, $C_n(x) = L(x)$.

This is an *infinite* (so-called *non-uniform*) computational model: for each input length n, there is a different "program" C_n for computing the 2^n inputs of that length, and the size of this program can grow with n.

Note that *every* language L has circuits of size $O(n2^n)$, following the observation that every $f : \{0,1\}^n \to \{0,1\}$ is specified by a 2^n-bit vector, called the *truth table* of f. This construction can be improved to $2^n/n + o(2^n/n)$ size [55, 71], and a simple counting argument shows that this improved size bound is tight for general functions. The class of functions of type (3.1) computable with "feasibly-sized" circuits is often called $\mathsf{P/poly}$:

Definition 3.2. Let $s : \mathbb{N} \to \mathbb{N}$ and $L : \{0,1\}^\star \to \{0,1\}$. Define $\mathsf{SIZE}(s(n))$ to be the class of functions L such that L has size-$s(n)$ circuits, and $\mathsf{P/poly}$ to be the class of functions L such that there is a $k \geq 1$ satisfying $L \in \mathsf{SIZE}(n^k + k)$.

Studying $\mathsf{P/poly}$ requires us to contemplate explicit trade-offs between the *sizes* of programs for computing functions and the *sizes* of inputs to those programs. Proving that a language is *not* in $\mathsf{P/poly}$ is a very strong result, implying that even finite segments of the language require "large" computations, relative to the sizes of inputs in the segment. From such results one can, in principle, derive concrete numerical statements about the limits of solving a problem. A proof that $L \notin \mathsf{P/poly}$ could potentially be used to establish that solving L on 1000-bit inputs requires 10^{100} size computations. This would be a true claim concerning the intractability of L in the known physical universe.[3]

Immediately one wonders how the two computational models of algorithms and circuits relate. The basic *Circuits for Algorithms* question is:

> What "normal" algorithms (efficient or not) can be simulated in $\mathsf{P/poly}$?

More precisely, take a complexity class $\mathcal{C}$ defined with respect to the usual uniform algorithm model (P, NP, PSPACE, EXP, NEXP, and so on). Which of these classes are contained in $\mathsf{P/poly}$? For example, if EXP were contained in $\mathsf{P/poly}$, then all uniform

[3] In fact, statements of this form have been extracted from circuit complexity lower bounds. See Stockmeyer-Meyer [74].

algorithms running in exponential time can be simulated by polynomial-size computations in the non-uniform circuit model. It is believed that in general, circuit families cannot really solve NP-hard problems significantly more efficiently than algorithms can, and that $\mathsf{NP} \not\subset \mathsf{P/poly}$. Complexity theory is *very* far from proving this; for one, it would imply $\mathsf{P} \neq \mathsf{NP}$.

To gain a little insight into the difficulty, we may first ask if P/poly is contained in any of the above classes. The answer to that question is no. Let $\{M_1, M_2, M_3, \ldots\}$ be a computable enumeration of Turing machines. Consider the function $L(x)$ defined to output 1 if and only if $M_{|x|}$ halts on $1^{|x|}$. For every n, either L outputs 1 on all n-bit strings, or L outputs 0 on all such strings. It is easy to infer from this that $L \in \mathsf{P/poly}$. However, L is also *undecidable*, as there is an easy reduction from the Halting Problem to L. The class P/poly, defined in terms of an infinite computational model, has unexpected power.

In general, the tools of computability theory are essentially powerless for understanding P/poly, and complexity theory has not yet discovered enough new tools. Indeed, this provides another reason to study circuit complexity: we're forced to develop new lower bound proof methods that go beyond old methods like diagonalization, which is known not to be sufficient by itself due to the Relativization barrier [9]. These new methods may be useful in the long run for resolving other problems such as P vs NP. While nontrivial results are known (which we now survey), they are meager in comparison to what is conjectured.

3.1. Classes with efficient circuits. It is relatively easy to see that $\mathsf{P} \subset \mathsf{P/poly}$: polynomial-time algorithms can be "unrolled" for polynomially many steps, and simulated step-by-step using polynomial-size circuits. Furthermore, randomized polynomial-time algorithms have polynomial-size circuit families, i.e., $\mathsf{BPP} \subset \mathsf{P/poly}$ [2], by judiciously hard-coding good random seeds in polynomial-size circuits.

Besides what we have already sketched, there are few other nontrivial results known. Kolmogorov made an intriguing conjecture:

Conjecture 3.3 (A. N. Kolmogorov, according to Levin [50]). *For every $L \in \mathsf{P}$, there is a k such that L has kn size circuits.*[4]

The conjecture would be surprising, if true. For languages in P requiring $n^{100^{100}}$ time, it appears unlikely that the complexity of such problems would magically shrink to $O(n)$ size, merely because a different circuit can be designed for each input length. Kolmogorov's conjecture implies $\mathsf{P} \neq \mathsf{NP}$ [50].

While it is generally believed that Conjecture 3.3 isn't true, a resolution looks very difficult. To see why, we sketch here the lack of progress on circuit lower bounds for languages in P. For a language $L : \{0,1\}^\star \to \{0,1\}$, define $L_n : \{0,1\}^n \to \{0,1\}$ to be *the n-bit restriction of L*: L_n agrees with L on all $x \in \{0,1\}^n$. The best known circuit lower bounds for functions in P are only small linear bounds:

Theorem 3.4 ([13]). *There is an $L \in \mathsf{P}$ with $C_{B_2}(L_n) \geq 3n - o(n)$ for all n.*

Theorem 3.5 ([38, 49]). *There is an $L \in \mathsf{P}$ with $C_{U_2}(L_n) \geq 5n - o(n)$ for all n.*

Hence it is possible that every $L \in \mathsf{P}$ has circuits of size $5.1n$. Even if the L is allowed to be in NP, no better circuit lower bounds are known. It is open whether every

[4] Apparently the conjecture was based on the affirmative answer by Kolmogorov and Arnol'd of Hilbert's 13th problem [7, 47], which asks if every continuous function on three variables can be expressed as a composition of finitely many continuous functions on two variables.

$L \in \mathsf{TIME}[2^{O(n)}]^{\mathsf{NP}}$ (functions in $2^{O(n)}$ time with access to an NP oracle) has $5.1n$ size circuits. In Section 5 we will see a possible approach to this question.

It was recently shown that, if Kolmogorov's conjecture is true, then such $O(n)$-size circuits must be intractable to construct algorithmically [67].[5]

3.2. Classes without efficient circuits. Let us now survey which functions *are* known to not be in P/poly.

Ehrenfeucht [22] studied the decision problem for sentences in the first order theory of $\mathbb{N}$ with addition, multiplication, and exponentiation, where all quantified variables are bounded by constants. (The problem is clearly decidable since all variables are bounded.) He showed that this problem requires $(1+\delta)^n$-size circuits for some $\delta > 0$, assuming a reasonable encoding of sentences as binary strings. Meyer (1972, cf. [74]) and Sholomov [72] proved that the same problem is decidable by a Turing machine using exponential ($2^{O(n)}$) space—in complexity notation, $\mathsf{EXPSPACE} \not\subset \mathsf{SIZE}((1+\delta)^n)$. This result can be scaled down to show the same circuit size lower bound for a language in $\Sigma_3\mathsf{EXP}$.[6]

Kannan [42] proved that $\mathsf{NEXP}^{\mathsf{NP}} \not\subset \mathsf{P}/\mathrm{poly}$. In fact his proof shows that $\mathsf{NEXP}^{\mathsf{NP}} \not\subset \mathsf{SIZE}(f(n))$, for every $f : \mathbb{N} \to \mathbb{N}$ satisfying $f(f(n)) \leq 2^n$ (these are the *half-exponential* functions). It is open whether $\mathsf{NEXP}^{\mathsf{NP}} \subset \mathsf{SIZE}(2^{\varepsilon n})$ for all $\varepsilon > 0$.

The P/poly lower bound of Kannan has been mildly improved over the years, to the presumably smaller (but still gigantic) complexity class MAEXP [15]. However, it is open whether NEXP (or even $\mathsf{EXP}^{\mathsf{NP}}$) is contained in P/poly. It looks impossible that all problems verifiable in *exponential time* could be computed using only polynomial-size circuits, but the infinite nature of the circuit model has confounded all proof attempts. Section 5 outlines a new approach to this problem.

3.3. Restricted circuits. There are several natural ways to restrict the circuit model beyond just circuit size, and still allow for complex circuit computations. In particular, restricting the depth leads to an array of possibilities.

Let A be the basis of *unbounded fan-in* AND and OR gates with NOT, i.e.,

$$A = \{\mathrm{NOT}\} \cup \bigcup_{n\in\mathbb{N}} \{\mathrm{OR}_n, \mathrm{AND}_n\}.$$

For an integer $m \geq 2$, let M_m be the basis of unbounded fan-in MODm, AND, and OR gates with NOT:

$$M_m = \{\mathrm{NOT}\} \cup \bigcup_{n\in\mathbb{N}} \{\mathrm{OR}_n, \mathrm{AND}_n, \mathrm{MOD}m_n\}.$$

Let T be the basis of unbounded fan-in MAJ gates with NOT:

$$T = \{\mathrm{NOT}\} \cup \bigcup_{n\in\mathbb{N}} \{\mathrm{MAJ}_n\}.$$

The following complexity classes are all subclasses of P/poly that have been widely studied.

[5] More formally, there is a language L computable in n^c time for some $c \geq 1$, such that for *every* $d \geq 1$ and every algorithm A running in n^d time, there are infinitely many n such that $A(1^n)$ does not output an $O(n)$ size circuit C_n computing L on n-bit inputs.

[6] $\Sigma_3\mathsf{EXP} = \mathsf{NEXP}^{\mathsf{NP}^{\mathsf{NP}}}$ is nondeterministic exponential time with oracle access to $\mathsf{NP}^{\mathsf{NP}}$ (and $\mathsf{NP}^{\mathsf{NP}}$ equals nondeterministic polynomial time with oracle access to NP). This class is contained in EXPSPACE, and the containment is probably proper.

Let $k \geq 0$ be an integer.

- NCk: Languages computable with polynomial size, $O(\log^k n)$ depth circuits over the basis U_2.[7]
- ACk: Languages computable with a polynomial size and $O(\log^k n)$ depth circuit family $\{C_n\}$ over A. That is, there is a fixed integer $d \geq 1$ such that every C_n has depth $d \log^k n$.[8]
- ACk[m]: Languages computable with polynomial size, $O(\log^k n)$ depth circuits over M_m.
- ACCk: The union over all $m \geq 2$ of ACk[m].[9]
- TCk: Languages computable with polynomial size, $O(\log^k n)$ depth circuits over the basis T.[10]

A thorough survey of these classes cannot be provided here; instead, let us focus attention on the most relevant aspects for the present story. The most well-studied of these classes are AC0, ACC0, TC0, and NC1, and it is known that

$$\mathsf{AC0} \subsetneq \mathsf{AC0[p]} \subsetneq \mathsf{ACC0} \subseteq \mathsf{TC0} \subseteq \mathsf{NC1} \subseteq \mathsf{P/poly},$$

when p is a prime power.

NC1 is well-motivated in several ways: for instance, it is also the class of languages computable with infinite families of polynomial-size Boolean *formulas*, or circuits where all internal gates have outdegree one. For formulas, interesting lower bounds are known: the best known formula size lower bound for a function in P is $n^{3-o(1)}$ over U_2, by Håstad [29]. TC0 is well-motivated from the study of neural networks: the MAJ function is a primitive model of a neuron, and the constant depth criterion reflects the massive parallelism of the human brain. Less primitive models of the neuron, such as *linear threshold functions*, end up defining the same class TC0. (A linear threshold function is a Boolean function f defined by a linear form $\sum_{i=1}^{n} w_i x_i$ for some $w_i \in \mathbb{Z}$, and a threshold value $t \in \mathbb{Z}$. For all $(x_1, \ldots, x_n) \in \{0,1\}^n$, $f(x_1, \ldots, x_n) = 1$ if and only if $\sum_i w_i x_i \geq t$.)

The MODm operations may look strange, but they arose naturally out of a specific program to develop circuit complexity in a "bottom up" way, starting with very restricted circuits and a hope of gradually relaxing the restrictions over time. First, AC0 was studied as a "maximally parallel" but still non-trivial class, and it was shown that MOD2 $\notin$ AC0 [3, 26]. This made it reasonable to ask what is computable when the MOD2 function is provided among the basis functions in AC0, leading to the definition of AC0[2]. Then it was proved that for distinct primes p and q, MODq $\notin$ AC0[q] [65, 73], hence MOD3 $\notin$ AC0[2]. One then wonders what is computable when MOD3 and MOD2 are both allowed in the basis. It is not hard to see that including MOD6 in the basis functions is equivalent to including MOD3 and MOD2. Attention turned to AC0[6]. (There were many other separate threads of research, such as lower bounds on fixed-depth versions of TC0 [28], which space prevents us from covering here.)

[7] The acronym NC stands for "Nick's Class," named after Nick Pippenger.

[8] AC stands for "Alternating Circuits," alternating between AND and OR. As a reminder, NOT gates are not counted in the depth bounds of AC, ACC, and TC circuits.

[9] ACC stands for "Alternating Circuits with Counting."

[10] TC stands for "Threshold Circuits."

At this point, the trail was lost. It is still open whether every language in P/poly (and in EXP) has depth-three circuit families over M_6. It has been shown only recently that NEXP is not contained in ACC0, via a generic connection between algorithms-for-circuits and circuits-for-algorithms [80, 82] (see Section 5). Yet it is open whether NEXP is contained in TC0, even for TC0 circuits of depth *three*.

4. Algorithms For circuits

In the most common form of circuit analysis problem, one takes a circuit as input, and decides a property of the function computed by the circuit. Let a property P be a function from the set of all Boolean functions $\{f : \{0,1\}^n \to \{0,1\} \mid n \geq 0\}$ to the set $\{0,1\}$.

Generic Circuit Analysis
Input: A logical circuit C
Output: A property $P(f)$ of the function f computed by C

The canonical example of such a problem is the Circuit Satisfiability problem (a.k.a. Circuit-SAT), which we shall survey in detail.

Circuit-SAT
Input: A logical circuit C
Output: Does the function f computed by C output 1 on some input?

This is basically equivalent to checking if C implements a trivial function that is constant on all inputs—a function of *minimum* circuit complexity. Hence the Circuit-SAT problem may viewed as providing nontrivial insight into the circuit complexity of the function implemented by a given circuit.

As Circuit-SAT is NP-complete, it is unlikely that there is an polynomial-time algorithm for it. An algorithm which exhaustively searches over all possible inputs to C requires $\Omega(2^n \cdot |C|)$ time steps, where n is the number of inputs to C, and $|C|$ is the size of the circuit. Is there a *slightly* faster algorithm, running in (for example) $1.99^n \cdot |C|^2$ time? Presently, there is no known algorithm for solving the problem on generic circuits of size s and n inputs that is asymptotically faster than the time cost of exhaustive search. Fine-grained questions of this variety are basic to two emerging areas of research: parameterized algorithms [21, 23] and exact algorithms [24]. For many NP-hard problems, asymptotically faster algorithms over exhaustive search do exist, and researchers actively study the extent to which exhaustive search can be beaten. (We shall see in Section 5 that even slightly faster Circuit-SAT algorithms can sometimes have a major impact.)

4.1. Restrictions of Circuit-SAT. As seen in Section 3, many circuit restrictions have been studied; here we survey the known algorithms for the satisfiability problem under these different restrictions. In this section, we think of AC0, ACC, TC0, NC1, and P/poly not as classes of languages, but as *classes of circuit families*: collections of infinite circuit families satisfying the appropriate restrictions. For each class $\mathcal{C}$, a satisfiability problem can be defined:

$\mathcal{C}$-SAT
Input: A circuit C from a family in class $\mathcal{C}$
Output: Is there an input on which C evaluates to true?

Just as with general Circuit-SAT, the $\mathcal{C}$-SAT problem remains NP-complete even for AC0-SAT [18], yet for simple enough $\mathcal{C}$, $\mathcal{C}$-SAT algorithms running faster than exhaustive search *are* known.

k-SAT. The k-SAT problem is to determine satisfiability of a very simple circuit type: an AND of ORs of k literals (which can be input variables and/or their negations). This is also called *conjunctive normal form* (CNF). Without loss of generality, the AND gate may be assumed to have $O(n^k)$ fan-in, as there are only $O(n^k)$ possible ORs of k literals. The k-SAT problem is also NP-complete [18] for all $k \geq 3$. Nevertheless, 3-SAT can be solved in 1.331^n time using a deterministic algorithm [56], or 1.308^n time [30] using a randomized algorithm. These running times form the tail end of a long line of published algorithms, with each subsequent algorithm decreasing the base of the exponent by a little bit. (See the survey of Dantsin and Hirsch [19].)

How much faster can 3-SAT be solved? The *Exponential Time Hypothesis* of Impagliazzo and Paturi [34] asserts that this line of work must "converge" to some base of exponent greater than 1:

Exponential Time Hypothesis (ETH): There is a $\delta > 0$ such that 3-SAT on n variables cannot be solved in $O((1+\delta)^n)$ time.

Impagliazzo, Paturi, and Zane [36] showed that ETH is not just a hypothesis about one NP-complete problem: by using clever *subexponential time reductions*, ETH implies that many other NP-hard problems require $(1+\delta)^n$ time to solve for some $\delta > 0$. Many other consequences of ETH have been found [51].

The k-SAT problem for arbitrary k has also been extensively studied. The best known k-SAT algorithms all run in $2^{n-n/(ck)}$ time, for a fixed constant c [19, 62, 63, 68]. So for $k > 3$, the savings in running time over 2^n slowly disappears as k increases. The *Strong Exponential Time Hypothesis* [16, 34] asserts that this phenomenon is inherent in all SAT algorithms:

Strong Exponential Time Hypothesis (SETH): For every $\delta < 1$ there is a k such that k-SAT on n variables cannot be solved in $2^{\delta n}$ time.

For example, SETH implies that even $2^{.99999n}$ is not enough time for solving k-SAT over all constants k. (It is known that SETH implies ETH.)

AC0-SAT. There has been less work on this problem, but recent years have seen progress [10, 16, 32]. The fastest known AC0-SAT algorithm is that of Impagliazzo, Matthews, and Paturi [32], who give an $O(2^{n-\Omega(n/(\log s)^{d-1})})$ time algorithm on circuits with n inputs, s gates, and depth d.

ACC0-SAT. The author [82] gave an algorithm running in $O(2^{n-n^{\varepsilon}})$ time for ACC0 circuits of $2^{n^{\varepsilon}}$ size, for some $\varepsilon \in (0,1)$ which is a function of the depth d of the given circuit and the modulus m used in the MODm gates. This algorithm was recently extended to handle the larger circuit class $\mathsf{ACC0} \circ \mathsf{THR}$, which is ACC0 augmented with an additional layer of arbitrary linear threshold gates near the inputs [81].

TC0-SAT. For depth-two TC0 circuits, Impagliazzo, Paturi, and Schneider [35] showed that satisfiability with n inputs and cn wires (i.e., edges) can be determined in $2^{\delta n}$ time for some $\delta < 1$ that depends on c. No nontrivial algorithms are known for satisfiability of depth-three TC0 (and circuit lower bounds aren't known, either).

Formula-SAT. Santhanam [66] proved that satisfiability of cn size formulas over U_2 can be determined in $2^{\delta n}$, for some $\delta < 1$ depending on c. His algorithm was extended to the basis B_2 by Seto and Tamaki [70], and to larger size formulas over U_2 by Chen et al. [17]. Applying recent concentration results of Komargodski, Raz and Tal [48], the algorithm of Chen et al. can solve SAT for formulas over U_2 of size $n^{3-o(1)}$ in randomized $2^{n-n^{\Omega(1)}}$ time with zero error. (Recall that the best known formula *lower bound* is $n^{3-o(1)}$ size as well; these Formula-SAT algorithms exploit similar ideas as in the lower bound methods.)

4.2. Approximate circuit analysis. A different form of circuit analysis is that of *additive approximate counting*; that is, approximating the *fraction of satisfying assignments* to a given circuit:

Circuit Approximation Probability Problem (CAPP)
Input: A circuit C
Output: The quantity $\Pr_x[C(x) = 1]$, to within $\pm\, 1/10$.

The constant $1/10$ is somewhat arbitrary, and could be any constant in $(0, 1/2)$ (usually this constant is a parameter in the algorithm). As with $\mathcal{C}$-SAT, the problem $\mathcal{C}$-CAPP can be defined for any circuit class $\mathcal{C}$. Approximate counting has been extensively studied due to its connections to *derandomization*. CAPP is easily computable with *randomness* by sampling (for instance) 100 x's uniformly at randomn, and evaluating C on them. We want to know purely *deterministic* algorithms. The structure of this subsection will parallel that of the coverage of $\mathcal{C}$-SAT. We cannot hope to cover all work in this article, and can only provide highlights.[11]

Several algorithms we shall mention give a stronger property than just approximately counting. Prior to viewing the circuit, these algorithms efficiently construct a small collection $\mathcal{A}$ of strings (assignments), such that for all circuits C of the appropriate size and depth from a circuit class $\mathcal{C}$, the fraction of satisfying assignments of C over $\mathcal{A}$ is a close approximation to the total fraction of satisfying assignments of C. Such algorithms are called *pseudorandom generators* and are inherently tied to lower bounds. Indeed, lower bounds against a circuit class $\mathcal{C}$ are generally a prerequisite for pseudorandom generators for $\mathcal{C}$, because the efficient process which produces such a collection $\mathcal{A}$ cannot be modeled within $\mathcal{C}$.

[11] We should also note that many algorithms from the previous subsection not only solve $\mathcal{C}$-SAT, but can *exactly* count the number of satisfying assignments (or can be modified to do so), implying a $\mathcal{C}$-CAPP algorithm.

The case of depth-two AC0 (i.e., of an AND of ORs of literals, or an OR of AND of literals) is especially interesting. Luby and Velickovic [53] showed that this case of CAPP is computable in $n^{\exp(O(\sqrt{\log\log n}))}$ time. Gopalan, Meka, and Reingold [27] improved this to about $n^{O(\log\log n)}$ time. It appears that here, a deterministic polynomial-time algorithm for CAPP may be within reach.

Ajtai and Wigderson [4] showed that AC0-CAPP is solvable in 2^{n^ε} time for every $\varepsilon > 0$, providing a pseudorandom generator. A pseudorandom generator of Nisan [59] yields an AC0-CAPP algorithm running in $n^{\log^{O(d)} s}$ time, where s is the size and d is the depth. There has been much work since then; most recently, Trevisan and Xue [76] construct tighter pseudorandom generators for AC0, showing that AC0-CAPP can be computed in $n^{\tilde{O}(\log^{d-1} s)}$ time.

For the class ACC0, *exact* counting of satisfying assignments can be done in about the same (best known) running time as computing satisfiability [81].

To our knowledge, no nontrivial CAPP algorithm for depth-two TC0 circuits is known. However, here is a good place to mention two other threads of work relating to low-depth circuits. The problem of approximately counting the number of zeroes in $\{0,1\}^n$ of a low-degree polynomial over a finite field is equivalent to computing CAPP on a MODp of AND gates of fan-in d. This problem can be solved essentially optimally for fixed d, in deterministic time $O_d(n^d)$ [14, 52, 54, 78]. A *polynomial threshold function of degree* d (PTF) has the form $f : \{-1,1\}^n \to \{-1,1\}$ and is representable by the sign of a multivariate degree-d polynomial over the integers. (Such functions can be construed as Boolean; the convention is that $-1, 1$ correspond to *true* and *false*, respectively.) Approximating the number of zeroes to a degree-d PTF can be modeled by solving CAPP on a *linear threshold gate* of MOD2 gates of fan-in d. It is known that for every fixed d, approximate counting for degree-d PTFs can be done in polynomial time [58].

For Boolean formulas, Impagliazzo, Meka, Zuckerman [33] give a pseudorandom generator yielding a $2^{s^{1/3+o(1)}}$ time algorithm for Formula-CAPP on size-s formulas over U_2. For formulas of size s over B_2 and branching programs of size s, their generator can be used to approximately count in $2^{s^{1/2+o(1)}}$ time.

No nontrivial results for CAPP are known for unrestricted Boolean circuits.

4.3. Truth table analysis. So far, we have only considered circuit analysis problems where the input to be analyzed is a circuit. Another class of circuit analysis problems take a *Boolean function on n variables* as input, specified as a 2^n-bit string, and the goal is to compute some property of "good" circuits which compute the function f.

Generic Truth Table Analysis
Input: A function $f : \{0,1\}^n \to \{0,1\}$
Output: Property $P(f)$ of circuits computing f

A natural example is that of minimizing a circuit given its truth table:

Circuit-Min [40, 83]
Input: A function $f : \{0,1\}^n \to \{0,1\}$ and $k \in \mathbb{Z}^+$
Output: Is $C_{B_2}(f) \leq k$?

In other words, we want to decide if the circuit complexity of f is at most k. As with Circuit-SAT and CAPP, we can also define the $\mathcal{C}$-Min problem for restricted circuit classes $\mathcal{C}$. The problem is easily seen to be in NP. It is strongly believed that Circuit-Min is intractable: if it were in P, then there would be no *pseudorandom functions*, contradicting conventional wisdom in cryptography. Informally, a pseudorandom function is a function f implementable with polynomial-size circuits that "behaves like" a random function, to all efficient processes with input/output access to f. Since a random function g has high circuit complexity with high probability, and f has low circuit complexity, an efficient algorithm for Circuit-Min could be used to tell f and g apart with non-negligible success probability, after querying them at $n^{O(1)}$ points. As a result, restricted versions of Circuit-Min such as NC1-Min and TC0-Min are also intractable under cryptographic assumptions, as those classes are believed to support such functions.[12]

Perhaps Circuit-Min is NP-hard. Proving that is a difficult open problem. To obtain a polynomial-time reduction from (say) 3-SAT to Circuit-Min, unsatisfiable formulas have to be efficiently mapped into functions without small circuits; however, recall that we do not *know* explicit functions with high circuit complexity. Kabanets and Cai [40] show that if the NP-hardness of Circuit-Min could be proved under a natural notion of reduction, then long-open circuit lower bounds like $\mathsf{EXP} \not\subset \mathsf{P/poly}$ would follow.

One version of Circuit-Min is known to be NP-complete: *DNF-Min*, the problem of minimizing a DNF formula (an OR of ANDs of literals) given its truth table [5, 57]. (Intuitively, DNF-Min can be proved hard because strong lower bounds are known for computing Boolean functions with DNFs.) However, one can efficiently find an *approximately* minimum-sized DNF [5].

A newly-introduced and related analysis problem is that of *compression*:

Compression of $\mathcal{C}$ [17]
Input: A function $f : \{0,1\}^n \to \{0,1\}$ computable with a circuit from $\mathcal{C}$
Output: A (possibly unrestricted) circuit C computing f with size $\ll 2^n/n$

Chen et al. [17] show that the techniques used in existing circuit lower bound proofs can be "mined" to obtain somewhat efficient compression algorithms for AC0, small Boolean formulas, and small branching programs. They pose as an open problem whether ACC0 admits such a compression algorithm.

Learning circuits There is one more important form of circuit analysis that can be viewed as restricted access to the truth table of a function: that of *learning* a function $f : \{0,1\}^n \to \{0,1\}$ which is initially hidden, but is known or assumed to be implementable in some restricted circuit class $\mathcal{C}$. In this survey we focus on the problem of *exact learning of* $\mathcal{C}$ *with membership and equivalence queries* [6], where a learning algorithm does not see f in its entirety, but has the ability to:

[12]Here is a good point to briefly mention a connection between Circuit-Min and complexity barriers. Razborov and Rudich [64] showed that practically all known circuit lower bound proof techniques (i.e., proving there are no efficient circuits-for-algorithms) yield weak *efficient* algorithms for Circuit-Min, weak enough to break any candidate pseudorandom function. Hence it's likely that such "natural proofs" cannot prove even TC0 lower bounds. In summary, every "natural proof" that there are no efficient circuits for some algorithms also yields an interesting algorithm for efficient circuits!

- query f on an arbitrary $x \in \{0,1\}^n$ (a *membership query*), and
- pose a hypothesis circuit H on n bits, asking if H and f compute the same function (an *equivalence query*). If $H \neq f$, the algorithm is provided with a counterexample point x on which $H(x) \neq f(x)$.

Pseudorandom functions, mentioned earlier, naturally connect with learning. A pseudorandom function has small circuits yet "looks like a random function" when it is queried a small number of times—this kind of function is naturally difficult to learn. Hence learning of Boolean functions computable in $\mathsf{TC0}$ and $\mathsf{NC1}$ is believed to be intractable. Other examples can be found in the references [44, 77].

5. Connections

In the *Circuits for Algorithms* space, one designs simple circuits to simulate complex algorithms, or proves that no simple circuits exist for this task. In *Algorithms for Circuits*, the goal is to design faster circuit-analysis algorithms. It is reasonable to hypothesize that these tasks may inform each other. A provably nontrivial algorithm for analyzing all circuits from a class should exhibit, at its core, nontrivial understanding about the limitations of that circuit class. Conversely, if a simple function cannot be computed by small circuits, then algorithms may be able to use this function to analyze small circuits faster than exhaustive search.

For restricted classes of circuits, one can sometimes adapt known techniques for proving lower bounds to derive faster SAT algorithms (or CAPP algorithms) for those circuits. For instance, the progress on Formula-SAT algorithms and on pseudorandom generators for Boolean formulas, both mentioned in Section 4, came out of tighter analyses of the *random restriction* method originally used for proving formula lower bounds [29, 75].

In the following, we restrict attention to more generic connections (i.e., formal implications) between efficient circuit-analysis algorithms and circuit lower bounds.

5.1. Circuit lower bounds and derandomization/CAPP. Perhaps the earliest explicit study of how algorithms and lower bounds connect can be found in the formal theory of cryptographic pseudorandomness, initiated by Blum and Micali [12] and Yao [84]. The existence of cryptographic pseudorandom generators were shown to imply subexponential time deterministic simulations of randomized polynomial time algorithms. Nisan and Wigderson [60] defined a relaxed notion of pseudorandom generator explicitly for the purposes of derandomizing randomized algorithms (instead of for cryptography) and proved connections between circuit lower bounds and the existence of pseudorandom generators. Subsequent work [8, 31, 37, 46] improved these connections. These papers give an effective *equivalence* between (for example) functions in $2^{O(n)}$ time requiring "high" circuit complexity, and the existence of pseudorandom generators computable in $2^{O(n)}$ time that are effective against "low complexity" circuits.

For an example, Babai et al. [8] showed that $\mathsf{EXP} \not\subset \mathsf{P/poly}$ implies that randomized polynomial-time algorithms can be simulated deterministically in subexponential time, on infinitely many input lengths. Formally speaking:

Theorem 5.1 ([8]). $\mathsf{EXP} \not\subset \mathsf{P/poly}$ *implies* $\mathsf{BPP} \subseteq \mathsf{ioSUBEXP}$.

This connection was sharpened by Impagliazzo and Wigderson:

Theorem 5.2 ([37]). *If there is a $\delta > 0$ and a function computable in $2^{O(n)}$ time requiring circuits of size at least $(1+\delta)^n$ for almost all input lengths n, then* $\mathsf{P} = \mathsf{BPP}$.

That is, from exponential-size lower bounds, one can simulate every randomized polynomial-time algorithm in deterministic polynomial time. Impagliazzo, Kabanets, and Wigderson [31] showed that even a seemingly weak lower bound like $\mathsf{NEXP} \not\subset \mathsf{P/poly}$ would imply a derandomization result: namely, there is a simulation of Merlin-Arthur games (a probabilistic version of NP) computable in nondeterministic subexponential time. In the opposite direction, they showed how a subexponential time algorithm for CAPP implies lower bounds:

Theorem 5.3 ([31]). *If CAPP can be computed in $2^{n^{o(1)}}$ time for all circuits of size n, then* $\mathsf{NEXP} \not\subset \mathsf{P/poly}$.

Recall the best known algorithm for CAPP is exhaustive search, taking $\Omega(2^n)$ time; an improvement to 2^{n^ε} for every $\varepsilon > 0$ would be an incredible achievement. However, the hypothesis of Theorem 5.3 can be weakened significantly: essentially any nontrivial improvement over 2^n time for CAPP implies the lower bound.

Theorem 5.4 ([80]). *Suppose for every k, CAPP on circuits of size n^k and n inputs can be computed in $O(2^n/n^k)$ time. Then* $\mathsf{NEXP} \not\subset \mathsf{P/poly}$.

Furthermore, computing CAPP for a restricted circuit class $\mathcal{C}$ faster than exhaustive search would imply that $\mathsf{NEXP} \not\subset \mathcal{C}$ [67, 80]. Theorem 5.4 requires that $\mathcal{C}$ satisfy certain closure properties (all classes covered in this survey satisfy them). Ben-Sasson and Viola [11] have recently sharpened the connection between CAPP algorithms and circuit lower bounds, by carefully modifying a known construction of probabilistically checkable proofs.

5.2. Circuit lower bounds from SAT algorithms. We now survey the impact of Circuit-SAT algorithms on the topic of Circuits for Algorithms. First, if we have "perfect" circuit analysis, i.e., Circuit-SAT is solvable in *polynomial time*, then there is a function in EXP that does not have small circuits. This result is quite old in complexity-theory years:

Theorem 5.5 (Meyer [43]). *If* $\mathsf{P} = \mathsf{NP}$ *then* $\mathsf{EXP} \not\subset \mathsf{P/poly}$.

This is an interesting implication, but it may be of limited utility since we do not believe the hypothesis. Nevertheless, Theorem 5.5 is a good starting point for thinking about how circuit analysis can relate to circuit lower bounds. A proof can be quickly sketched: assuming $\mathsf{P} = \mathsf{NP}$, we obtain many other equalities between complexity classes, including $\mathsf{NP}^{\mathsf{NP}^{\mathsf{NP}}} = \mathsf{P}$ and $\Sigma_3\mathsf{EXP} = \mathsf{NEXP}^{\mathsf{NP}^{\mathsf{NP}}} = \mathsf{EXP}$. As stated in Section 3, $\Sigma_3\mathsf{EXP}$ contains a language requiring circuits of maximum complexity (by directly "diagonalizing" against all circuits up to the maximum size). Therefore EXP now contains such a language as well.

This simple argument shows how a feasibility hypothesis like $\mathsf{P} = \mathsf{NP}$ implies a reduction in the algorithmic complexity of hard functions. It is tantalizing to wonder if a lower bound could proved by contradiction, in this way: from a feasibility hypothesis, deduce that the complexity of another *provably hard* function reduces so drastically that it becomes contradictorily *easy*. Sure enough, recent progress by the author on ACC0 lower bounds (described below) takes this approach.

Studying the proof more carefully, Theorem 5.5 can be improved in a few ways. Considering the contrapositive of the proof sketch, we find that if every function in $2^{O(n)}$ time has less than the *maximum possible* circuit complexity $(1 + o(1))2^n/n$, then $\mathsf{P} \neq \mathsf{NP}$. In other words, if non-uniform circuits can gain even a small advantage over exponential-time algorithms in simulation, then $\mathsf{P} \neq \mathsf{NP}$ would follow. Another improvement of Theorem 5.5 comes from observing we do not exactly need *polynomial time* Circuit-SAT algorithms: weaker guarantees such as $n^{(\log n)^k}$ time would suffice to conclude $\mathsf{EXP} \not\subset \mathsf{P/poly}$. Assuming ETH, this sort of running time is still beyond what is expected.

Combining these results with our earlier remarks on derandomization, we see that either EXP doesn't have large circuits and hence $\mathsf{P} \neq \mathsf{NP}$, or EXP requires large circuits and every randomized algorithm would have an interesting deterministic simulation, by Theorem 5.2. No matter how EXP vs $\mathsf{P/poly}$ is resolved, the consequences will be very interesting.

Modern times. Theorem 5.5 and its offshoots only work for Circuit-SAT algorithms running in subexponential time. An indication that techniques for weak SAT algorithms may still be useful for circuit lower bounds appears in the work of Paturi, Pudlak, and Zane [63]. They gave a structure lemma on k-SAT instances, and applied it to prove not only that k-SAT has an $2^{n-n/k}$ time algorithm, but also lower bounds for depth-three $\mathsf{AC0}$ circuits.

In recent years, the author showed that very weak improvements over exhaustive search for $\mathcal{C}$-SAT would imply circuit lower bounds for NEXP:

Theorem 5.6 ([80, 82]). *There is a $c > 0$ such that, if $\mathcal{C}$-SAT can be solved on circuits with n inputs and n^k size in $O(2^n/n^c)$ time for every k, then $\mathsf{NEXP} \not\subset \mathcal{C}$.*

While the conclusion is weaker than Theorem 5.5, the hypothesis (for all classes $\mathcal{C}$ we have considered) is *extremely* weak compared to $\mathsf{P} = \mathsf{NP}$; indeed, it even looks plausible. The above theorem was combined with the $\mathsf{ACC0}$-SAT algorithm mentioned in Section 4.1 to conclude:

Theorem 5.7 ([82]). $\mathsf{NEXP} \not\subset \mathsf{ACC0}$.

Since Theorem 5.7 was proved, it has been concretely extended twice. The first extension slightly lowers the complexity of NEXP, down to complexity classes such as $\mathsf{NEXP/1} \cap \mathsf{coNEXP/1}$ [79]. (In fact a generic connection is proved between $\mathcal{C}$-SAT algorithms and $\mathcal{C}$ circuit lower bounds for $\mathsf{NEXP/1} \cap \mathsf{coNEXP/1}$, with a slightly stronger hypothesis: we have to assume SAT algorithms for $n^{\log^k n}$ size circuits.) The second extension strengthens $\mathsf{ACC0}$ up to the class $\mathsf{ACC0} \circ \mathsf{THR}$, or $\mathsf{ACC0}$ circuits augmented with a layer of linear threshold gates near the inputs [81].

Theorem 5.6 holds for all circuit classes $\mathcal{C}$ of Section 2, but one may need (for example) a SAT algorithm for $2d$-depth circuits to obtain a d-depth circuit lower bound. The project of tightening parameters to make $\mathcal{C}$-SAT algorithms directly correspond to the same $\mathcal{C}$ circuit lower bounds has seen much progress [11, 39, 61, 67]. Now (for example) it is known that SAT algorithms for depth $d + 1$ or $d + 2$ (depending on the gate basis) imply depth-d lower bounds.

Perhaps Circuit-SAT looks too daunting to improve upon. Are there other connections between SAT algorithms and circuit lower bounds? Yes. From faster 3-SAT algorithms, superlinear size lower bounds follow:

Theorem 5.8 ([80]). *Suppose the Exponential Time Hypothesis (ETH) is false: that is, 3-SAT is in $2^{\varepsilon n}$ time for every $\varepsilon > 0$. Then there is a language $L \in \mathsf{TIME}[2^{O(n)}]^{\mathsf{NP}}$ such that, for every $c \geq 1$, L does not have cn-size circuits.*

ETH was discussed in Section 4.1, and the conclusion of Theorem 5.8 was discussed as open in Section 3.1. Refuting the Strong Exponential Time Hypothesis (SETH) from Section 4.1 also implies (weaker) circuit lower bounds:

Theorem 5.9 ([39]). *Suppose SETH is false: that is, there is a $\delta < 1$ such that k-SAT is in $O(2^{\delta n})$ time for all k. Then there is a language $L \in \mathsf{TIME}[2^{O(n)}]^{\mathsf{NP}}$ such that, for every $c \geq 1$, L does not have cn-size Valiant-series-parallel circuits.*

Intuition for the connections. One intuition is that a faster circuit-analysis algorithm (say, for $\mathcal{C}$-SAT) demonstrates a specific *weakness* in representing computations with circuits from $\mathcal{C}$. A circuit family from $\mathcal{C}$ is *not* like a collection of black boxes which can easily hide satisfying inputs. (If we could only query the circuit as a black box, viewing only its input/output behavior, we could not solve $\mathcal{C}$-SAT in $o(2^n)$ time.) Another intuition is that the existence of a faster circuit-analysis algorithm for $\mathcal{C}$ demonstrates a *strength* of algorithms that run in less-than-2^n time: they can analyze nontrivial properties of a given circuit. Hence from assuming a less-than-2^n time $\mathcal{C}$-SAT algorithm, we should be capable of inferring that "less-than-2^n time algorithms are strong" and "$\mathcal{C}$-circuits are weak."

These observations hint at a proof that, assuming a $\mathcal{C}$-SAT algorithm, there is a language in NEXP without polynomial-size $\mathcal{C}$ circuits. The actual proof does not resemble these hints; it is a proof by contradiction. We assert that both a faster algorithm for analyzing $\mathcal{C}$ exists, and that $\mathsf{NEXP} \subset \mathcal{C}$. Together these two assumptions imply a too-good-to-be-true algorithm: a way to simulate every language solvable in nondeterministic $O(2^n)$ time with only $o(2^n)$ time. This simulation contradicts the *nondeterministic time hierarchy theorem* [85], which implies that there are problems solvable in 2^n time nondeterministically which cannot be solved in $O(2^n/n)$ time nondeterministically. Informally, the faster nondeterministic simulation works by using $\mathsf{NEXP} \subset \mathcal{C}$ to nondeterministically guess $\mathcal{C}$ circuits that help perform an arbitrary $2^{O(n)}$ time computation, and using the faster circuit-analysis algorithm to verify that these $\mathcal{C}$ circuits do the job correctly.

5.3. Other connections.

Circuit lower bounds from learning. Intuitively, an efficient algorithm for learning circuits would have to harness some deep properties about the circuit class under consideration; perhaps these properties would also be enough to prove circuit lower bounds. Fortnow and Klivans proved a theorem modeling this intuition. Let $\mathcal{C}$ be a restricted circuit class, such as those defined in Section 3.3. In the following, say that $\mathcal{C}$ is *exactly learnable* if there is an algorithm for learning every hidden function from $\mathcal{C}$ using membership and equivalence queries (cf. Section 4.3).

Theorem 5.10 ([25]). *If all n-bit functions from $\mathcal{C}$ are exactly learnable in deterministic $2^{n^{o(1)}}$ time, then $\mathsf{EXP}^{\mathsf{NP}} \not\subset \mathcal{C}$.*

Theorem 5.11 ([25]). *If all n-bit functions from $\mathcal{C}$ are exactly learnable in randomized polynomial time, then randomized exponential time (BPEXP) is not contained in $\mathcal{C}$.*

Recently, these connections between learning circuits and circuit lower bounds have been somewhat strengthened:

Theorem 5.12 ([45]). *If $\mathcal{C}$ is exactly learnable in $2^{n^{o(1)}}$ time, then there is a language in $\mathsf{TIME}[2^{n^{o(1)}}]$ that is not in $\mathcal{C}$.*

Theorem 5.13 ([45]). *If $\mathcal{C}$ is exactly learnable in polynomial time, then there is a language in $\mathsf{TIME}[n^{\omega(1)}]$ that is not in $\mathcal{C}$.*

These proofs use a clever diagonalization argument, where the learning algorithm is used to construct an efficiently computable function f that plays the role of a *contrarian teacher* for the learning algorithm. When the learner asks a membership query x, f tells the learner *true* if f has not already committed to a value for x (otherwise, f reports $f(x)$). When an equivalence query is asked, f tells the learner "not equivalent" and outputs the first string y for which it has not already committed to an output value (thereby committing to a value for y). As f is constructed to never be equivalent to any hypothesis proposed by the learning algorithm, f cannot have circuits in $\mathcal{C}$.

Equivalences between circuit analysis and circuit lower bounds. Earlier it was mentioned that there are rough equivalences between pseudorandom generators and circuit lower bounds. Pseudorandom generators can be viewed as "circuit analysis" algorithms, in the context of computing CAPP. Impagliazzo, Kabanets, and Wigderson [31] proved an explicit equivalence:

Theorem 5.14 ([31]). $\mathsf{NEXP} \not\subset \mathsf{P}/poly$ *if and only if CAPP is in* $\mathsf{ioNTIME}[2^{n^{\varepsilon}}]/n^{\varepsilon}$ *for all* $\varepsilon > 0$.

Without going into the notation, this theorem states that NEXP circuit lower bounds are equivalent to the existence of "non-trivial" subexponential time algorithms for CAPP. The author recently proved a related equivalence between the $\mathsf{NEXP} \not\subset \mathcal{C}$ problem (for various circuit classes $\mathcal{C}$) and circuit-analysis algorithms. Call an algorithm A *non-trivial for $\mathcal{C}$-Min* if

- $A(f)$ runs in $2^{O(n)}$ time on a given $f : \{0,1\}^n \to \{0,1\}$, and
- for all constants k and for infinitely many input lengths n, there is a $f : \{0,1\}^n \to \{0,1\}$ such that $A(f)$ outputs 1, and for all $f : \{0,1\}^n \to \{0,1\}$ computable with an $(n^k + k)$-size circuit from $\mathcal{C}$, $A(f)$ outputs 0.

That is, for infinitely many n, algorithm A outputs 1 on at least one Boolean function on n bits, and 0 on all functions with small circuit complexity.

Theorem 5.15 ([79]). $\mathsf{NEXP} \not\subset \mathcal{C}$ *if and only if there is an algorithm A which is non-trivial for $\mathcal{C}$-Min.*

Connections in an algebraic setting. In this survey, we considered Boolean functions and circuits computing them. However, connections between circuit-analysis algorithms and circuit lower bounds also hold in an *algebraic* framework, where Boolean functions are replaced by *polynomials over a ring R*, and Boolean circuits are replaced by *algebraic circuits*, which defined analogously to Boolean circuits, but we allow side constants from the ring as

extra inputs to an algebraic circuit, and the gates are either *additions* or *multiplications* over the ring. Typically, R is taken to be a finite field, or $\mathbb{Z}$. Each algebraic circuit $C(x_1, \ldots, x_n)$ computes some polynomial $p(x_1, \ldots, x_n)$ over R.

The canonical circuit-analysis problem in this setting is:

Polynomial Identity Testing (PIT): Given an algebraic circuit C, does C compute the identically zero polynomial?

Using subtraction, it is easy to see this problem is equivalent to determining if two algebraic circuits C and C' compute the *same* polynomial.

It's natural to think of PIT as a type of satisfiability problem. However, PIT is probably *not* NP-hard: the problem is easily solvable in *randomized polynomial time* by substituting random elements (possibly over an extension field) [20, 69, 86]. A very interesting open problem is to determine whether randomness is necessary for efficiently solving PIT. Kabanets and Impagliazzo [41] proved that an efficient *deterministic* algorithm for PIT would imply *algebraic circuit lower bounds*: either $\mathsf{NEXP} \not\subset \mathsf{P/poly}$, or the permanent of a matrix requires superpolynomial-size algebraic circuits.

6. Conclusion

This article has shown how a host of open problems in algorithms have direct bearing on some of the central problems in complexity theory. It is quite likely that there exist deeper interactions between Algorithms for Circuits and Circuits for Algorithms which await our discovery. Hopefully, the reader has been persuaded to think a little more about how algorithms and lower bounds relate to each other.

Acknowledgements. This article is an adaptation of an invited paper in the proceedings of the 2014 IEEE Conference on Computational Complexity. The author thanks Boaz Barak, Sam Buss, Lance Fortnow, Valentine Kabanets, and Rahul Santhanam for comments on an earlier version. The author was supported by the Alfred P. Sloan foundation, Microsoft Research, and the NSF under grant CCF-1212372.

References

[1] Scott Aaronson and Avi Wigderson, *Algebrization: A new barrier in complexity theory*, ACM TOCT, **1**, 2009.

[2] Leonard Adleman, *Two theorems on random polynomial time*, In FOCS, 1978, pp. 75–83.

[3] Miklos Ajtai, Σ_1^1*-formulae on finite structures*, Annals of Pure and Applied Logic, **24** (1983), 1–48.

[4] Miklós Ajtai and Avi Wigderson, *Deterministic simulation of probabilistic constant depth circuits (preliminary version)*, In FOCS, 1985, pp. 11–19.

[5] Eric Allender, Lisa Hellerstein, Paul McCabe, Toniann Pitassi, and Michael Saks, *Minimizing DNF formulas and AC0 circuits given a truth table*, SIAM J. Comput., **38**(1) (2008), 63–84.

[6] Dana Angluin, *Queries and concept learning*, Machine Learning, **2**(4) (1987), 319–342.

[7] V. I. Arnol'd, *On functions of three variables*, Dokl. Akad. Nauk SSSR, **114** (1957), 679–681.

[8] László Babai, Lance Fortnow, Noam Nisan, and Avi Wigderson, *BPP has subexponential time simulations unless EXPTIME has publishable proofs*, Computational Complexity, **3**(4) (1993), 307–318.

[9] Theodore Baker, John Gill, and Robert Solovay, *Relativizations of the P =? NP question*, SIAM J. Comput., **4**(4) (1975), 431–442.

[10] Paul Beame, Russell Impagliazzo, and Srikanth Srinivasan, *Approximating AC0 by small height decision trees and a deterministic algorithm for # AC0SAT*, In CCC, 2012, pp. 117–125.

[11] Eli Ben-Sasson and Emanuele Viola, *Short* PCPs *with projection queries*, In ICALP, 2014, page to appear.

[12] Manuel Blum and Silvio Micali, *How to generate cryptographically strong sequence of pseudo-random bits*, SIAM J. Comput., **13** (1984), 850–864.

[13] Norbert Blum, *A boolean function requiring 3n network size*, Theoretical Computer Science, **28** (1984), 337–345.

[14] Andrej Bogdanov and Emanuele Viola, *Pseudorandom bits for polynomials*, SIAM J. Comput., **39**(6) (2010), 2464–2486.

[15] Harry Buhrman, Lance Fortnow, and Thomas Thierauf, *Nonrelativizing separations*, In CCC, 1998, pp. 8–12.

[16] Chris Calabro, Russell Impagliazzo, and Ramamohan Paturi, *The complexity of satisfiability of small depth circuits*, In Parameterized and Exact Complexity (IWPEC), 2009, pp. 75–85.

[17] Ruiwen Chen, Valentine Kabanets, Antonina Kolokolova, Ronen Shaltiel, and David Zuckerman, *Mining circuit lower bound proofs for meta-algorithms*, In CCC, 2014, page to appear.

[18] Stephen Cook, *The complexity of theorem-proving procedures*, In STOC, 1971, pp. 151–158.

[19] Evgeny Dantsin and Edward A. Hirsch, *Worst-case upper bounds*, In Handbook of Satisfiability. Frontiers in Artificial Intelligence and Applications, volume 185, IOS Press, 2009, pp. 403–424.

[20] Richard A. DeMillo and Richard J. Lipton, *A probabilistic remark on algebraic program testing*, Information Processing Letters, **7**(4) (1978), 192–195.

[21] Rodney G. Downey and Michael R. Fellows, *Parameterized Complexity*, Springer-Verlag, 1999.

[22] Andrzej Ehrenfeucht, *Practical decidability*, J. Comput. Syst. Sci., **11** (1975), 392–396.

[23] Jörg Flum and Martin Grohe, *Parameterized complexity theory*, Springer Heidelberg, 2006.

[24] Fedor V. Fomin and Dieter Kratsch, *Exact Exponential Algorithms*, Springer, 2010.

[25] Lance Fortnow and Adam R. Klivans, *Efficient learning algorithms yield circuit lower bounds*, J. Comput. Syst. Sci., **75**(1) (2009).

[26] Merrick Furst, James Saxe, and Michael Sipser, *Parity, circuits, and the polynomial-time hierarchy*, Mathematical Systems Theory, **17**(1) (April 1984), 13–27. See also FOCS'81.

[27] Parikshit Gopalan, Raghu Meka, and Omer Reingold, *Dnf sparsification and a faster deterministic counting algorithm*, Computational Complexity, **22**(2) (2013), 275–310.

[28] András Hajnal, Wolfgang Maass, Pavel Pudlák, Mario Szegedy, and György Turán, *Threshold circuits of bounded depth*, J. Comput. Syst. Sci., **46**(2) (1993), 129–154.

[29] Johan Håstad, *The shrinkage exponent of de morgan formulae is 2*, SIAM J. Comput., **27** (1998), 48–64.

[30] Timon Hertli, *3-SAT faster and simpler - Unique-SAT bounds for PPSZ hold in general*, In FOCS, 2011, pp. 277–284.

[31] Russell Impagliazzo, Valentine Kabanets, and Avi Wigderson, *In search of an easy witness: Exponential time vs. probabilistic polynomial time*, J. Comput. Syst. Sci., **65**(4) (2002), 672–694.

[32] Russell Impagliazzo, William Matthews, and Ramamohan Paturi, *A satisfiability algorithm for* AC^0, In SODA, 2012, pp. 961–972.

[33] Russell Impagliazzo, Raghu Meka, and David Zuckerman, *Pseudorandomness from shrinkage*, In FOCS, 2012, pp. 111–119.

[34] Russell Impagliazzo and Ramamohan Paturi, *On the complexity of k-SAT*, J. Comput. Syst. Sci., **62**(2) (2001), 367–375.

[35] Russell Impagliazzo, Ramamohan Paturi, and Stefan Schneider, *A satisfiability algorithm for sparse depth two threshold circuits*, In FOCS, 2013, pp. 479–488.

[36] Russell Impagliazzo, Ramamohan Paturi, and Francis Zane, *Which problems have strongly exponential complexity?*, J. Comput. Syst. Sci., **63**(4) (2001), 512–530.

[37] Russell Impagliazzo and Avi Wigderson, *P = BPP if E requires exponential circuits: Derandomizing the XOR lemma*, In STOC, 1997, pp. 220–229.

[38] Kazuo Iwama and Hiroki Morizumi, *An explicit lower bound of 5n - o(n) for boolean circuits*, In MFCS, 2002, pp. 353–364.

[39] Hamidreza Jahanjou, Eric Miles, and Emanuele Viola, *Local reductions*, Technical Report TR13-099, Electronic Colloquium on Computational Complexity, July 2013.

[40] Valentine Kabanets and Jin-Yi Cai, *Circuit minimization problem*, In STOC, 2000, pp. 73–79.

[41] Valentine Kabanets and Russell Impagliazzo, *Derandomizing polynomial identity tests means proving circuit lower bounds*, Computational Complexity, **13**(1-2) (2004), 1–46.

[42] Ravi Kannan, *Circuit-size lower bounds and non-reducibility to sparse sets*, Information and Control, **55**(1) (1982), 40–56.

[43] Richard Karp and Richard Lipton, *Turing machines that take advice*, L'Enseignement Mathématique, **28**(2) (1982), 191–209.

[44] Michael J. Kearns and Leslie G. Valiant, *Cryptographic limitations on learning boolean formulae and finite automata*, JACM, **41**(1) (1994), 67–95.

[45] Adam Klivans, Pravesh Kothari, and Igor C. Oliveira, *Constructing hard functions using learning algorithms*, In CCC, 2013 pp. 86–97.

[46] Adam Klivans and Dieter van Melkebeek, *Graph nonisomorphism has subexponential size proofs unless the polynomial hierarchy collapses*, SIAM J. Comput., **31**(5) (2002), 1501–1526.

[47] A. N. Kolmogorov, *On the representation of continuous functions of several variables by superposition of continuous functions of a smaller number of variables*, Dokl. Akad. Nauk SSSR, **108** (1956), 179–182.

[48] Ilan Komargodski, Ran Raz, and Avishay Tal, *Improved average-case lower bounds for DeMorgan formulas*, In FOCS, 2013, pp. 588–597.

[49] Oded Lachish and Ran Raz, *Explicit lower bound of 4.5n - o(n) for boolean circuits*, In STOC, 2001, pp. 399–408.

[50] Richard Lipton, *Some consequences of our failure to prove non-linear lower bounds on explicit functions*, In Structure in Complexity Theory Conference, 1994, pp. 79–87.

[51] Daniel Lokshtanov, Dániel Marx, and Saket Saurabh, *Lower bounds based on the exponential time hypothesis*, Bulletin of the EATCS, **105** (2011), 41–72.

[52] Shachar Lovett, *Unconditional pseudorandom generators for low degree polynomials*, Theory of Computing, **5**(1) (2009), 69–82.

[53] Michael Luby and Boban Velickovic, *On deterministic approximation of DNF*, Algorithmica, **16**(4/5) (1996) 415–433.

[54] Michael Luby, Boban Velickovic, and Avi Wigderson, *Deterministic approximate counting of depth-2 circuits*, In Proceedings of the 2nd ISTCS, 1993, pp. 18–24.

[55] O. B. Lupanov, *A method of circuit synthesis*, Izvestiya VUZ, Radiofizika, **1**(1) (1959), 120–140.

[56] Kazuhisa Makino, Suguru Tamaki, and Masaki Yamamoto, *Derandomizing* HSSW *algorithm for 3-sat*, In COCOON, 2011, pp. 1–12.

[57] W. J. Masek, *Some NP-complete set covering problems*, Manuscript, 1979.

[58] Raghu Meka and David Zuckerman, *Pseudorandom generators for polynomial threshold functions*, SIAM J. Comput., **42**(3) (2013), 1275–1301.

[59] Noam Nisan, *Pseudorandom bits for constant depth circuits*, Combinatorica, **11**(1) (1991), 63–70.

[60] Noam Nisan and Avi Wigderson, *Hardness vs randomness*, J. Comput. Syst. Sci., **49**(2) (1994), 149–167.

[61] Igor Oliveira, *Algorithms versus circuit lower bounds*, Technical Report TR13-117, ECCC, September 2013.

[62] Ramamohan Paturi, Pavel Pudlák, Michael E. Saks, and Francis Zane, *An improved exponential-time algorithm for* k-SAT, JACM, **52**(3) (2005), 337–364. (See also FOCS'98.)

[63] Ramamohan Paturi, Pavel Pudlák, and Francis Zane, *Satisfiability coding lemma*, Chicago J. Theor. Comput. Sci., 1999, 1999. See also FOCS'97.

[64] Alexander Razborov and Steven Rudich, *Natural proofs*, J. Comput. Syst. Sci., **55**(1) (1997), 24–35.

[65] Alexander A. Razborov, *Lower bounds on the size of bounded-depth networks over the complete basis with logical addition*, Mathematical Notes of the Academy of Sciences of the USSR, **41**(4) (1987), 333–338.

[66] Rahul Santhanam, *Fighting perebor: New and improved algorithms for formula and qbf satisfiability*, In FOCS, 2010, pp. 183–192.

[67] Rahul Santhanam and Ryan Williams, *On medium-uniformity and circuit lower bounds*, In CCC, 2013, pp. 15–23.

[68] Uwe Schöning, *A probabilistic algorithm for k-SAT based on limited local search and restart*, Algorithmica, **32**(4) (2002), 615–623.

[69] Jacob Schwartz, *Fast probabilistic algorithms for verification of polynomial identities*, JACM, **27**(4) (1980), 701–717.

[70] Kazuhisa Seto and Suguru Tamaki, *A satisfiability algorithm and average-case hardness for formulas over the full binary basis*, Computational Complexity, *22*(2) (2013), 245–274. See also CCC'12.

[71] Claude E. Shannon, *The synthesis of two-terminal switching circuits*, Bell Syst. Techn. J., **28** (1949), 59–98.

[72] L. A. Sholomov, *On one sequence of functions which is hard to compute*, Mat. Zametki, **17** (1975), 957–966.

[73] Roman Smolensky, *Algebraic methods in the theory of lower bounds for Boolean circuit complexity*, In STOC, 1987, pp. 77–82.

[74] Larry J. Stockmeyer and Albert R. Meyer, *Cosmological lower bound on the circuit complexity of a small problem in logic*, JACM, **49**(6) (2002), 753–784.

[75] B. A. Subbotovskaya, *Realizations of linear functions by formulas using +,*,-*, Soviet Mathematics Doklady, **2** (1961), 110–112.

[76] Luca Trevisan and TongKe Xue, *A derandomized switching lemma and an improved derandomization of AC0*, In CCC, 2013, pp. 242–247.

[77] Leslie G. Valiant, *A theory of the learnable*, In STOC, 1984, pp. 436–445.

[78] Emanuele Viola, *The sum of d small-bias generators fools polynomials of degree d*, Computational Complexity, **18**(2) (2009), 209–217.

[79] Ryan Williams, *Natural proofs versus derandomization*, In STOC, 2013, pp. 21–30.

[80] Ryan Williams, *Improving exhaustive search implies superpolynomial lower bounds*, SIAM J. Comput., **42**(3) (2013), 1218–1244. See also STOC'10.

[81] Ryan Williams, *New algorithms and lower bounds for circuits with linear threshold gates*, In STOC, 2014, page to appear.

[82] Ryan Williams, *Nonuniform ACC circuit lower bounds*, JACM, **61**(1) (2014), 2. See also CCC'11.

[83] S. V. Yablonksi, *The algorithmic difficulties of synthesizing minimal switching circuits*, Dokl. Akad. Nauk SSSR, **124**(1) (1959), 44–47.

[84] Andrew Yao, *Theory and application of trapdoor functions*, In FOCS, 1982, pp. 80–91.

[85] Stanislav Žák, *A Turing machine time hierarchy*, Theoretical Computer Science, *26*(3) (1983), 327–333.

[86] R. E. Zippel, *Probabilistic algorithms for sparse polynomials*, In International Symposium on Symbolic and Algebraic Manipulation, 1979, pp. 216–226.

Computer Science Department, Stanford University, Stanford, CA, USA
E-mail: rrw@cs.stanford.edu

Codes with local decoding procedures

Sergey Yekhanin

Abstract. Error correcting codes allow senders to add redundancy to messages, encoding bit strings representing messages into longer bit strings called codewords, in a way that the message can still be recovered even if a fraction of the codeword bits are corrupted. In certain settings however the receiver might not be interested in recovering all the message, but rather seek to quickly recover just a few coordinates of it. Codes that allow one to recover individual message coordinates extremely fast (locally), from accessing just a small number of carefully chosen coordinates of a corrupted codeword are said to admit a local decoding procedure. Such codes have recently played an important role in several areas of theoretical computer science and have also been used in practice to provide reliability in large distributed storage systems. We survey what is known about these codes.

Mathematics Subject Classification (2010). Primary 94B05, 94B35; Secondary 68R05, 68P20.

Keywords. Error correcting codes, locally decodable codes, private information retrieval schemes, multiplicity codes, matching vectors codes, local reconstruction codes, maximally recoverable codes.

1. Introduction

The 60+ years of research in coding theory that started with the works of Shannon [24] and Hamming [14] gave us nearly optimal ways to add redundancy to messages, encoding bit strings representing messages into longer bit strings called codewords, in a way that the message can still be recovered even if a certain fraction of the codeword bits are corrupted.

In certain scenarios however, the receiver of the corrupted message might not be interested in recovering all the message, but rather seek to reconstruct just a small portion of it. For instance one can think of a setting where a large database is stored encoded with an error correcting code and a user who is willing to access a single database record. When classical codes are employed the user would have no alternative but to decode all the database (investing effort that is at least proportional to the database size) and then access the record. This example calls for codes that admit local decoding procedures, i.e., allow one to reliably recover individual message coordinates from accessing just a small number of coordinates of a corrupted codeword. The goal of our survey is to review the state of the art in such codes.

In what follows we model corrupted coordinates as being erased rather then flipped. This simplifies presentation and also allows us give a unified treatment of the few lines of work on the subject. We assume that the decoder is aware of which coordinates are missing. Below is a simple example of a code that admits a local decoding procedure.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Hadamard code. The code encodes k-bit messages $\mathbf{x}$ to 2^k-bit codewords $C(\mathbf{x})$. It allows any coordinate $\mathbf{x}(i)$ to be recovered by accessing just two coordinates of $C(\mathbf{x})$ even after almost a half of coordinates of $C(\mathbf{x})$ are erased. In what follows, let $[k]$ denote the set $\{1, \ldots, k\}$. Every coordinate of the Hadamard code corresponds to one (of 2^k) subsets of $[k]$ and stores the XOR of the corresponding bits of the message $\mathbf{x}$. Observe that for any set $S \subseteq [k]$, $\mathbf{x}(i)$ equals the XOR of the values stored at coordinates S and $S \triangle \{i\}$. (Here, $\triangle$ denotes the symmetric difference of sets such as $\{1, 4, 5\} \triangle \{4\} = \{1, 5\}$, and $\{1, 4, 5\} \triangle \{2\} = \{1, 2, 4, 5\}$). It is not difficult to verify that if less than half of coordinates of $C(\mathbf{x})$ are erased; then for every $i \in [k]$, there exists a set $S \subseteq [k]$ such that both coordinates corresponding to S and $S \triangle \{i\}$ are available, and thus $\mathbf{x}(i)$ can be recovered with two reads.

Codes with local decoding procedures vary in terms of the number of erasures after which local recovery can be guaranteed. The other two main parameters of interest are the codeword length and the query complexity. The length of the code measures the amount of redundancy that is introduced into the message by the encoder. The query complexity counts the number of coordinates that need to be read from the corrupted codeword in order to recover a single coordinate of the message. For instance, in the Hadamard code above, local recovery is guaranteed after $2^{k-1} - 1$ erasures, redundancy equals $2^k - k$, and query complexity is 2.

In general one cannot optimize all three parameters discussed above simultaneously. There are tradeoffs. In this survey we restrict our attention to two main families of codes with local decoding procedures, namely, Locally Decodable Codes (LDCs) and Local Reconstruction Codes (LRCs). Locally decodable codes allow quick recovery of individual message coordinates in a very aggressive scenario when a linearly growing number of codeword coordinates might be missing. These codes play an important role in several areas of theoretical computer science and tend to require either a large amount of redundancy or a high query complexity. Local reconstruction codes, by contrast, only allow quick recovery when just a single coordinate is unavailable. As such they are considerably more efficient in terms of both codeword length and number of queries. Instances of these codes have been used in practice to provide reliability in large distributed storage systems.

Outline of the paper. In Sections 2 through 4 we deal with locally decodable codes. Section 2 provides a basic introduction to the area and discusses applications. Sections 3 and 4 cover the two main families of LDCs, namely multiplicity codes that are the most efficient codes in the regime of high query complexity and matching vector codes that are the best known codes in the regime of low query complexity. In Section 5 we review the state of the art in local reconstruction codes.

2. Basics of locally decodable codes

As we discussed above LDCs are erasure correcting codes that allow extremely efficient (sub-linear time) decoding of individual message coordinates even when a linearly growing number of codeword coordinates are unavailable. Below is a formal definition.

Definition 2.1. Let q be a prime power, $1 \leq r \leq k \leq N$ be integers, and $\delta > 0$ be real. Assume that for every $i \in [k]$ there is a collection $\mathcal{D}_i$ of r-subsets of $[N]$. An (r, δ)-locally

decodable code is a mapping $C : \mathbb{F}_q^k \to \mathbb{F}_q^N$ such that:

- For every message $\mathbf{x} \in \mathbb{F}_q^k$, for each $i \in [k]$ and $S \in \mathcal{D}_i$ the symbol $\mathbf{x}(i)$ can be recovered from accessing r coordinates of $C(\mathbf{x})$ that belong to S.
- For every set $E \subseteq [N]$ such that $|E| \leq \delta n$, for every $i \in [k]$ there exists a set $S \in \mathcal{D}_i$ such that $E \cap S = \emptyset$.

A locally decodable code is called *linear* if C is a linear transformation over $\mathbb{F}_q$. Almost all codes considered in the survey are linear.

Not all parameters of locally decodable codes are considered equally important. In what follows we will typically pay little attention to alphabet size q and fraction of erasures δ and focus on the values of the codeword length N and the query complexity r when the message length k grows to infinity. Ideally, one would like to have both N and r as small as possible. One however can not minimize the length and the query complexity simultaneously. There is a tradeoff. On one end of the spectrum we have classical error correcting codes that have both query complexity and codeword length proportional to the message length. On the other end we have the Hadamard code that has query complexity 2 and codeword length exponential in the message length. Establishing the optimal trade-off between the length and the query complexity is the major goal of research in the area of locally decodable codes.

2.1. Reed Muller codes. In this section we discuss the oldest and most basic family of locally decodable codes. An LDC allows to quickly recover any coordinate of a message by accessing only few coordinates of its corrupted encoding. A related property is that of local correctability allowing to locally recover not only coordinates of the message but also arbitrary coordinates of the encoding.

Definition 2.2. Let q be a prime power, $1 \leq r \leq k \leq N$ be integers, and $\delta > 0$ be real. Assume that for every $i \in [N]$ there is a collection $\mathcal{D}_i$ of r-subsets of $[N]$. An (r, δ)-locally correctable code is a subset $C \subseteq \mathbb{F}_q^N$ of size q^k such that:

- For every codeword $\mathbf{x} \in C$, for each $i \in [N]$ and $S \in \mathcal{D}_i$ the symbol $\mathbf{x}(i)$ can be recovered from accessing r coordinates of $\mathbf{x}$ that belong to S.
- For every set $E \subseteq [N]$ such that $|E| \leq \delta n$, for every $i \in [N]$ there exists a set $S \in \mathcal{D}_i$ such that $E \cap S = \emptyset$.

We often refer to the quantity $\log_q |C|$ as the message length of a locally correctable code C. It is not hard to show that every linear locally correctable code yields a linear locally decodable code with the same parameters.

Reed Muller codes that we discuss below are locally correctable. In what follows a D-evaluation of a function h defined over a domain D, is a vector of values of h at all points of D. Also with a slight abuse of terminology we often refer to a dimension N of a vector $\mathbf{x} \in \mathbb{F}_q^N$ as its *length*. The key idea behind Reed Muller codes is that of polynomial interpolation. Messages are encoded by complete evaluations of low degree multivariate polynomials over a finite field. Local correctability is achieved through reliance on the rich structure of short local dependencies between such evaluations at multiple points.

A Reed Muller code is specified by three integer parameters, namely, a prime power (alphabet size) q, number of variables n, and a degree d. The q-ary code consists of $\mathbb{F}_q^n$-evaluations of all polynomials of total degree at most d in the ring $\mathbb{F}_q[z_1, \ldots, z_n]$. When

viewed as an LDC such code encodes $k = \binom{n+d}{d}$-long messages over $\mathbb{F}_q$ to q^n-long codewords.

Below we present the simplest local corrector for Reed Muller codes. To recover the value of a degree d polynomial $F \in \mathbb{F}_q[z_1, \ldots, z_n]$ at a point $\mathbf{w} \in \mathbb{F}_q^n$ it picks an affine line through $\mathbf{w}$ and then relies on the local dependency between the values of F at any $d+2$ points along the line. Let $\mathbb{F}_q^*$ denote the multiplicative subgroup of the field $\mathbb{F}_q$.

Theorem 2.3. *Let n and d be positive integers. Let q be a prime power, $\delta > 0$ be a real, and $d < (1-\delta)q - 1$ be an integer; then there exists a linear code of dimension $k = \binom{n+d}{d}$ in $\mathbb{F}_q^N$, $N = q^n$, that is $(d+1, \delta)$-locally correctable.*

Proof. The code consists of $\mathbb{F}_q^n$-evaluations of all polynomials of total degree at most d in the ring $\mathbb{F}_q[z_1, \ldots, z_n]$. Consider the i-th coordinate, $i \in [N]$ corresponding to a point $\mathbf{w} \in \mathbb{F}_q^n$. The family $\mathcal{D}_i$ consists of all $(d+1)$-tuples of points that can be obtained by picking a nontrivial affine line

$$L = \{\mathbf{w} + \lambda \mathbf{v} \mid \lambda \in \mathbb{F}_q^*\} \tag{2.1}$$

through $\mathbf{w}$ and fixing some $d+1$ points on it. The local correction procedure is quite natural. The decoder reads the values of the polynomial F at $d+1$ points of some undamaged set $S \in \mathcal{D}_i$. Note that such a set always exists under the assumptions of the theorem. Assume the set S comes from line (2.1). The decoder invokes univariate polynomial interpolation to recover the degree d polynomial f which is the restriction of F to the line L, i.e., $f(\lambda) = F(\mathbf{w} + \lambda \mathbf{v})$. The decoder outputs $f(0) = F(\mathbf{w})$. □

The method behind Reed Muller codes is simple and general. It yields codes for all possible values of query complexity r, i.e., one can set r to be an arbitrary function of the message length k by specifying an appropriate relation between n and d and letting these parameters grow to infinity. Increasing d relative to n yields shorter codes of larger query complexity. Below we summarize asymptotic parameters of several families of locally decodable codes based on Reed Muller codes.

r	N
$O(1)$	$\exp\left(k^{1/(r-1)}\right)$
$(\log k)^t, t > 1$	$k^{1+1/(t-1)+o(1)}$
$O(k^{1/t} \log^{1-1/t} k), t \geq 1$	$t^{t+o(t)} \cdot k$

2.2. Applications. Interestingly, the natural application of locally decodable codes to data storage mentioned in Section 1 is neither the historically earliest nor the most important. LDCs have a host of applications in other areas of theoretical computer science such as complexity theory, data structures, and derandomization. However their most prominent application is in cryptography to the design of Private Information Retrieval schemes (PIRs). In what follows we briefly review this application.

Private information retrieval schemes are cryptographic protocols designed to safeguard the privacy of database users. They allow clients to retrieve records from replicated public databases while completely hiding the identity of the retrieved records from database owners. In such protocols, users query each server holding the database. The protocol ensures that each individual server (by observing only the query it receives) gets no information about the identity of the items of user's interest. Below we demonstrate a general procedure that

obtains an r-server PIR scheme out of any r-query *smooth* LDC. A locally decodable code is called smooth if for every $i \in [k]$, the r-tuples in $\mathcal{D}_i$ cover the universe $[N]$ uniformly, i.e., each $j \in [N]$ belongs to the same number of sets in $\mathcal{D}_i$. Almost all known constructions of LDCs yield smooth codes.

Let C be a smooth LDC encoding k-bit messages to N-bit codewords. At the preprocessing stage servers $\mathcal{S}_1, \ldots, S_r$ encode the k-bit database $\mathbf{x}$ with the code C. Next the user $\mathcal{U}$ who is interested in obtaining the i-th bit of $\mathbf{x}$, picks an r-tuple of queries $(\mathrm{que}_1, \ldots, \mathrm{que}_r) \in \mathcal{D}_i$ uniformly at random. For every $j \in [r]$, the user sends the query que_j to the server $\mathcal{S}_j$. Each server $\mathcal{S}_j$ responds with a one bit answer $C(\mathbf{x})_{\mathrm{que}_j}$, which is the value of $C(\mathbf{x})$ at coordinate que_j. The user combines servers' responses to obtain $\mathbf{x}(i)$.

It is straightforward to verify that the protocol above is private since by the smoothness property for every $j \in [r]$ the query que_j is uniformly distributed over the set of codeword coordinates. The total communication is given by

$$r(\log N + 1).$$

Thus short codes of low query complexity yield communication efficient PIR schemes involving a small number of servers. For example, instantiating the reduction above with the best known 3-query LDCs yields 3-server private information retrieval schemes with $O\left(2^{\sqrt{\log k \log\log k}}\right)$ communication to access a k-bit database.

2.3. Notes. Formal definition of locally decodable codes was given in 2000 by Katz and Trevisan [18], who cited Leonid Levin for inspiration. See also [25]. However codes that meet this definition have been around for much longer. The first such family of codes, namely, Reed Muller codes [21, 23], were introduced in 1950s, and their local decodability properties have been exploited implicitly since then. Today there are few families of locally decodable codes that surpass Reed Muller codes in terms of query complexity vs. codeword length tradeoff. We are going to discuss two of them (namely, multiplicity codes [20] and matching vector codes [9, 28]) in the following sections. For a detailed survey of the locally decodable codes see [29]. Private Information Retrieval (PIR) schemes were introduced in [4]. See [27] for a recent survey.

3. Multiplicty codes

When dealing with codes that have low redundancy it is convenient to utilize the notion of code rate. For an (r, δ)-locally decodable code $C : \mathbb{F}_q^k \to \mathbb{F}_q^N$, the rate k/N is simply the ratio of the number of message symbols to the number of codeword symbols. Similarly, for a (r, δ)-locally correctable code $C \subseteq \mathbb{F}_q^N$, the rate is the ratio $(\log_q |C|)/N$. In applications to data transmission and storage one is naturally interested in codes of high rate, i.e., rate close to 1.

In this section we review multiplicity codes. These codes generalize Reed Muller codes (Theorem 2.3) and improve upon them in the high rate regime. Observe that with Reed Muller codes the rate can never be too high. Recall that these codes are specified by three parameters: alphabet size q, total degree d, and the number of variables n. The rate is highest when $n = 2$, $d = (1 - \delta)q$, and q grows to infinity. In this setting we have $k = \binom{d+2}{2}$ and $N = q^2$, thus the rate is $\binom{d+2}{2}/q^2 \approx \frac{1}{2}$ and query complexity $r = O(\sqrt{k})$.

Note that with Reed Muller codes, one cannot increase the rate by simply allowing evaluations of higher degree polynomials, as if one allows the degree to exceed the field size, one starts getting polynomials with colliding evaluations such as z and z^q. Multiplicity codes, however, use much higher degree polynomials and thus have significantly improved rates, and avoid the pitfall mentioned above by evaluating polynomials *together with their partial derivatives*.

In what follows we review the construction of multiplicity codes. We consider the simplest example of these codes based on bivariate polynomials, which have improved rate above $\frac{1}{2}$, and see how to locally correct them with essentially the same query complexity $O(\sqrt{k})$. Finally, we mention how general multiplicity codes are defined and discuss some of the ideas that go into locally correcting them. Our main result gives codes that simultaneously have rate approaching one, and allow for local correction with arbitrary polynomially-small number of queries.

3.1. Bivariate multiplicity codes. Let q be a prime power, let $\delta > 0$ and let integer $d = 2(1-\delta)q$. The multiplicity code of *order two* evaluations of degree d bivariate polynomials over $\mathbb{F}_q$ is the code defined as follows. As before, the coordinates are indexed by $\mathbb{F}_q^2$ (so $N = q^2$) and the codewords are indexed by bivariate polynomials of degree at most d over $\mathbb{F}_q$. However the alphabet now is $\mathbb{F}_q^3$. The codeword corresponding the polynomial $F(x_1, x_2)$ is the vector

$$C(F) = \left\langle \left(F(\mathbf{w}), \frac{\partial F}{\partial x_1}(\mathbf{w}), \frac{\partial F}{\partial x_2}(\mathbf{w}) \right) \right\rangle_{\mathbf{w} \in \mathbb{F}_q^2} \in \left(\mathbb{F}_q^3\right)^{q^2}.$$

In words, the $\mathbf{w}$ coordinate consists of the evaluation of F and its formal partial derivatives $\frac{\partial F}{\partial x_1}$ and $\frac{\partial F}{\partial x_2}$ at $\mathbf{w}$. Because two distinct polynomials of degree at most d can agree with multiplicity two on at most $d/2q$-fraction of the points in $\mathbb{F}_q^2$ no two codewords defined above collide. Since the alphabet size is now q^3, the rate of the new code is

$$\frac{\log_{q^3} q^{\binom{d+2}{2}}}{q^2} = \frac{\binom{d+2}{2}}{3q^2} \approx \frac{2}{3}(1-\delta)^2.$$

Summarizing the differences between this multiplicity code with the Reed Muller code described earlier:

- Instead of polynomials of degree $(1-\delta)q$, we consider polynomials of degree double of that.
- Instead of evaluating the polynomials, we take their *order two* evaluation.

This yields a code with the rate improved from below $1/2$ to nearly $2/3$. We now argue that the new code is still locally correctable with $O(\sqrt{k})$ queries.

Local correction of multiplicity codes: Given the codeword corresponding to the polynomial $F(x_1, x_2)$ with some (say, $\delta/2$) fraction of coordinates erased and given a point $\mathbf{w} \in \mathbb{F}_q^2$, we want to recover the symbol at coordinate $\mathbf{w}$, namely

$$\left(F(\mathbf{w}), \frac{\partial F}{\partial x_1}(\mathbf{w}), \frac{\partial F}{\partial x_2}(\mathbf{w}) \right). \tag{3.1}$$

Similarly to the case of Reed Muller codes, the algorithm picks a direction $\mathbf{v} \in \mathbb{F}_q^2$ such that less than a δ fraction of coordinates in the affine line

$$L = \{\mathbf{w} + \lambda\mathbf{v} \mid \lambda \in \mathbb{F}_q\}$$

are missing. Most directions are like this. Our intermediate goal is to recover the univariate polynomial $f(\lambda) = F(\mathbf{w} + \lambda\mathbf{v})$. The important observation here is that for every $\lambda_0 \in \mathbb{F}_q$, the $\mathbf{w}+\lambda_0\mathbf{v}$ coordinate of $C(F)$ completely determines both the value and the first derivative of the univariate polynomial $f(\lambda)$ at the point λ_0. Indeed,

$$\begin{aligned} f(\lambda_0) &= F(\mathbf{w} + \mathbf{v}\lambda_0), \\ \frac{\partial f}{\partial \lambda}(\lambda_0) &= \frac{\partial F}{\partial x_1}(\mathbf{w} + \mathbf{v}\lambda_0) \cdot \mathbf{v}(1) + \frac{\partial F}{\partial x_2}(\mathbf{w} + \mathbf{v}\lambda_0) \cdot \mathbf{v}(2), \end{aligned}$$

where the last identity follows by the chain rule. Thus our knowledge of $C(F)$ at $(1-\delta)q+1$ locations on the line L gives us access to $(1-\delta)q+1$ evaluations of the polynomial $f(\lambda)$ and its derivative $\frac{\partial f}{\partial \lambda}(\lambda_0)$, where $f(\lambda)$ is of degree $\leq 2(1-\delta)q$. This is enough to recover the polynomial $f(\lambda)$. Evaluating $f(\lambda)$ at $\lambda = 0$ gives us $F(\mathbf{w})$. Evaluating the derivative $\frac{\partial f}{\partial \lambda}(\lambda)$ at $\lambda = 0$ gives us the directional derivative of F at $\mathbf{w}$ in the direction $\mathbf{v}$ (which equals $\frac{\partial F}{\partial x_1}(\mathbf{w}) \cdot \mathbf{v}(1) + \frac{\partial F}{\partial x_2}(\mathbf{w}) \cdot \mathbf{v}(2)$).

We have clearly progressed towards our goal of computing $C(F)_{\mathbf{w}}$ given by formula (3.1), but we are not yet there. The final observation is that if we pick another direction $\mathbf{v}'$, that is not collinear with $\mathbf{v}$ and repeat the above process to recover the directional derivative of F at $\mathbf{w}$ in direction $\mathbf{v}'$, then the two directional derivatives of F at $\mathbf{w}$ in directions $\mathbf{v}, \mathbf{v}'$ together suffice to recover $\frac{\partial F}{\partial x_1}(\mathbf{w})$ and $\frac{\partial F}{\partial x_2}(\mathbf{w})$, as desired. This algorithm makes less than $2q$ queries, which is $O(\sqrt{k})$.

3.2. General multiplicity codes. The basic example of a multiplicity code above already achieves rate above $1/2$. To get codes of rate approaching 1, one needs to modify the construction by considering evaluations of all derivatives of F up to an even higher order. In order to locally recover the higher-order derivatives of F at a point $\mathbf{w}$, the decoding algorithm picks many lines passing through $\mathbf{w}$, recovers the restriction of F to those lines, and combines all these recovered univariate polynomials in a certain way.

To reduce the query complexity to $O(k^\epsilon)$ for small ϵ, one needs to modify the above example by considering multivariate polynomials in a larger number of variables n. The local decoding algorithm for this case, in order to locally recover at a point $\mathbf{w} \in \mathbb{F}_q^n$, still decodes by picking lines passing through $\mathbf{w}$; the reduced query complexity occurs because lines (with only q points) are now much smaller relative to a higher dimensional space $\mathbb{F}_q^n$.

Increasing both the maximum order of derivatives taken and the number of variables simultaneously yields multiplicity codes with rate close to one and arbitrarily low polynomial query complexity.

Theorem 3.1. *Let q be an arbitrary prime power. For every real $\epsilon, \alpha > 0$ there exists a real $\delta > 0$ such that for all sufficiently large message lengths k, there exists an $\mathbb{F}_q$-linear $(O(k^\epsilon), \delta)$-locally correctable code of rate $1 - \alpha$.*

3.3. Notes. Multiplicity codes were introduced in [20]. The construction builds on some technical tools from [7]. Alternative constructions with similar parameters have been given in [13, 15]. It is plausible that the query complexity of locally decodable codes of rate close to one can be further reduced. The only available lower bound is $\Omega(\log k)$ from [18].

4. Matching vector codes

In this section we review locally decodable codes that arise from families of matching vectors. Matching Vector (MV) codes are important as they exhibit dramatically better parameters than Reed Muller codes in the regime of low query complexity. Any construction of such codes naturally falls into two parts: the design of a matching vector family, and the actual code construction. Our focus is on the second part.

4.1. The framework. MV codes inherit some structure from Reed Muller codes. A matching vector code consists of a linear subspace of polynomials in $\mathbb{F}_q[z_1, \ldots, z_n]$, evaluated at all points of $\mathbb{C}_m^n$, where $\mathbb{C}_m$ is a certain multiplicative subgroup of $\mathbb{F}_q^*$. The decoding algorithm is similar to local decoders for Reed Muller codes. It operates by picking a line in a certain direction and decoding along it. The difference is that the monomials which are used are not of low degree, they are chosen according to a matching family of vectors. Further, the lines for decoding are *multiplicative*, a notion that we define shortly. In what follows let $\mathbb{Z}_m$ denote the ring of integers modulo an integer m. Also, let $\mathbf{u} \cdot \mathbf{v}$ denote the usual dot product of vectors $\mathbf{u}$ and $\mathbf{v}$.

Definition 4.1. Let $S \subseteq \mathbb{Z}_m \setminus \{0\}$. We say that families $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ and $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ of vectors in $\mathbb{Z}_m^n$ form an S-matching family if the following two conditions are satisfied:

- For all $i \in [k]$, $\mathbf{u}_i \cdot \mathbf{v}_i = 0$;
- For all $i, j \in [k]$ such that $i \neq j$, $\mathbf{u}_j \cdot \mathbf{v}_i \in S$.

We now show how one can obtain a matching vector locally decodable code out of a matching family. We start with some notation.

- We assume that q is a prime power, m divides $q-1$, and denote the unique subgroup of $\mathbb{F}_q^*$ of order m by $\mathbb{C}_m$;
- We fix some generator g of $\mathbb{C}_m$;
- For $\mathbf{w} \in \mathbb{Z}_m^n$, we define $g^{\mathbf{w}} \in \mathbb{C}_m^n$ by $\left(g^{\mathbf{w}(1)}, \ldots, g^{\mathbf{w}(n)}\right)$;
- For $\mathbf{w}, \mathbf{v} \in \mathbb{Z}_m^n$ we define the multiplicative line $M_{\mathbf{w},\mathbf{v}}$ through $\mathbf{w}$ in direction $\mathbf{v}$ to be the multi-set
$$M_{\mathbf{w},\mathbf{v}} = \left\{g^{\mathbf{w}+\lambda\mathbf{v}} \mid \lambda \in \mathbb{Z}_m\right\}; \tag{4.1}$$
- For $\mathbf{u} \in \mathbb{Z}_m^n$, we define the monomial
$$\mathrm{mon}_{\mathbf{u}} \in \mathbb{F}_q[z_1, \ldots, z_n]/(z_1^m = 1, \ldots, z_n^m = 1)$$
by
$$\mathrm{mon}_{\mathbf{u}}(z_1, \ldots, z_n) = \prod_{\ell \in [n]} z_\ell^{\mathbf{u}(\ell)}. \tag{4.2}$$

Observe that for any $\mathbf{w}, \mathbf{u}, \mathbf{v} \in \mathbb{Z}_m^n$ and $\lambda \in \mathbb{Z}_m$ we have

$$\mathrm{mon}_{\mathbf{u}}\left(g^{\mathbf{w}+\lambda\mathbf{v}}\right) = g^{\mathbf{u}\cdot\mathbf{w}}\left(g^{\lambda}\right)^{\mathbf{u}\cdot\mathbf{v}}. \tag{4.3}$$

This suggests that if we set $y = g^{\lambda} \in \mathbb{F}_q^*$ in formula (4.3); then what we get is a univariate polynomial in y. Hence the $M_{\mathbf{w},\mathbf{v}}$-evaluation of a monomial $\mathrm{mon}_{\mathbf{u}}$ is a $\mathbb{C}_m$-evaluation of a univariate monomial

$$g^{\mathbf{u}\cdot\mathbf{w}} y^{\mathbf{u}\cdot\mathbf{v}} \in \mathbb{F}_q[y]. \tag{4.4}$$

This observation is the foundation of all decoding algorithms for MV codes.

We now specify the encoding procedure and the most basic decoding procedure. Let $\mathcal{U}, \mathcal{V}$ be an S-matching family in $\mathbb{Z}_m^n$, where $|\mathcal{U}| = |\mathcal{V}| = k$.

Encoding: We encode a message $(\mathbf{x}(1), \ldots, \mathbf{x}(k)) \in \mathbb{F}_q^k$ by the $\mathbb{C}_m^n$-evaluation of the polynomial

$$F(z_1, \ldots, z_n) = \sum_{j=1}^{k} \mathbf{x}(j) \cdot \mathrm{mon}_{\mathbf{u}_j}(z_1, \ldots, z_n). \tag{4.5}$$

Notice that $F = F_{\mathbf{x}}$ is a function of the message $\mathbf{x}$ (we will omit the subscript and treat $\mathbf{x}$ as fixed throughout this section).

Decoding: The input to the decoder is an $\mathbb{C}_m^n$-evaluation of F with some δ fraction of coordinates erased and an index $i \in [k]$.

1. The decoder picks $\mathbf{w} \in \mathbb{Z}_m^n$ such that none of the values of F at points of the the multiplicative line $M_{\mathbf{w},\mathbf{v}_i}$ are erased. If $\delta < \frac{1}{m}$ such a $\mathbf{w}$ exists.
2. The decoder recovers the noiseless restriction of F to $M_{\mathbf{w},\mathbf{v}_i}$. To accomplish this the decoder queries the $M_{\mathbf{w},\mathbf{v}_i}$-evaluation of F at $|S| + 1$ locations

$$\left\{ g^{\mathbf{w}+\lambda\mathbf{v}_i} \mid \lambda \in \{0, \ldots, s\} \right\}. \tag{4.6}$$

Firstly, let us see how the $M_{\mathbf{w},\mathbf{v}_i}$-evaluation of F uniquely determines $\mathbf{x}(i)$. Observe that by formulas (4.3), (4.4) and (4.5) the $M_{\mathbf{w},\mathbf{v}_i}$-evaluation of F is the $\mathbb{C}_m$-evaluation of a polynomial

$$f(y) = \sum_{j=1}^{k} \mathbf{x}(j) \cdot g^{\mathbf{u}_j\cdot\mathbf{w}} y^{\mathbf{u}_j\cdot\mathbf{v}_i} \in \mathbb{F}_q[y]. \tag{4.7}$$

Properties of the S-matching family $\mathcal{U}, \mathcal{V}$ imply that $y^{(\mathbf{u}_j,\mathbf{v}_i)} = 1$, if $j = i$; and $y^{\mathbf{u}_j\cdot\mathbf{v}_i} = y^s$, for some $s \in S$ otherwise. Formula (4.7) yields

$$f(y) = \mathbf{x}(i) \cdot g^{\mathbf{u}_i\cdot\mathbf{w}} + \sum_{s\in S} \left(\sum_{j\,:\,\mathbf{u}_j\cdot\mathbf{v}_i = s} \mathbf{x}(j) \cdot g^{\mathbf{u}_j\cdot\mathbf{w}} \right) y^s. \tag{4.8}$$

For a polynomial $h \in \mathbb{F}_q[y]$ we denote by $\mathrm{supp}(h)$ the set of monomials with non-zero coefficients in h, where a monomial y^e is identified with the integer e. It is evident from formula (4.8) that $\mathrm{supp}(f) \subseteq S \cup \{0\}$ and

$$\mathbf{x}(i) = f(0)/g^{\mathbf{u}_i\cdot\mathbf{w}}. \tag{4.9}$$

Secondly, let us note that recovering the polynomial (4.8) from the values $\{c_0, \ldots, c_s\}$ of F at locations locations (4.6) is quite straightforward. The decoder simply recovers the

unique sparse univariate polynomial $h(y) \in \mathbb{F}_q[y]$ with $\mathrm{supp}(h) \subseteq S \cup \{0\}$ such that for all $\lambda \in \{0, \ldots, s\}$, $h(g^\lambda) = c_\lambda$. The uniqueness of such $h(y) = f(y)$ follows from standard properties of Vandermonde matrices.

Putting it all together we obtain the following

Proposition 4.2. *Let $\mathcal{U}, \mathcal{V}$ be a family of S-matching vectors in $\mathbb{Z}_m^n$, $|\mathcal{U}| = |\mathcal{V}| = k$, $|S| = s$. Suppose $m \mid q-1$, where q is a prime power; then there exists a $\mathbb{F}_q$-linear code encoding k-long messages to m^n-long codewords that is $(s+1, \frac{1}{m})$-locally decodable.*

As Proposition 4.2 suggests parameters of matching vector codes are governed by parameters of the underlying family of matching vectors. To get short codes of low query complexity we need large S-matching families for small sets S. The best constructions of such families are given by the following

Proposition 4.3. *Let $m = p_1 \ldots p_t$ be a product of t distinct primes. There exists a set $S \subseteq \mathbb{Z}_m \setminus \{0\}$, $|S| = 2^t - 1$ such that for all sufficiently large integers n, there is an S-matching family in $\mathbb{Z}_m^n$ of size*

$$n^{c\left(\frac{\log n}{\log \log n}\right)^{t-1}},$$

where the constant c depends only on m.

Combing the two propositions above we get

Theorem 4.4. *Let $m = p_1 \ldots p_t$ be a product of t distinct primes. Let q be a prime power such that $m \mid q-1$; then for infinitely many values of message length k there exists an $\mathbb{F}_q$-linear $(2^t, \frac{1}{m})$-locally decodable code of codeword length*

$$N = \exp\exp\left(O(\sqrt[t]{\log k (\log\log k)^{t-1}})\right).$$

Observe that for constant t the function above grows slower than any exponential function of the form $2^{\alpha k}$ though faster than any polynomial k^c.

4.2. Notes. Constructions of locally decodable codes from matching vectors originated in [28] and were developed further in [2, 6, 9]. An important progress in this line of work has been accomplished by Klim Efremenko in [9] where the first constructions of codes from matching vectors modulo composites (rather than primes) were considered. Proposition 4.3 is due to Grolmusz [12]. An important ingredient to his proof is the low-degree representation of the OR-function from [1].

Despite considerable progress in constructions of locally decodable codes of small query complexity we are still very far from closing the gap to lower bounds. It is only in the setting of 2-query codes that we know the true codeword length of optimal LDC, which is exponential [19]. For any other number of queries large gaps remain. For instance, in the case of the three-query codes the best upper bound for the codeword length comes from matching vector codes and is $\exp\exp(\sqrt{\log k \log\log k})$ while the best lower bound is $\Omega(k^2)$ from [26]. Closing this gap is a major open problem.

Locally decodable codes are also of interest over infinite fields. Questions about these codes relate to classical problems in combinatorial geometry [5].

5. Local reconstruction codes

In previous sections we reviewed the state of the art in locally decodable codes, i.e., codes that admit local recovery of individual message symbols in the regime when a linearly growing number of codeword coordinates may be unavailable. As we saw these codes are either highly redundant or have a large query complexity. In particular, to get rate close to one the best known LDCs need polynomially many queries, and to get constant query complexity independent of the message length they need super-polynomial codeword length.

In the current section we turn our attention to local reconstruction codes which only allow local recovery when just a single coordinate is unavailable, while also providing non-local recovery guarantees after a larger number of erasures. Since LRCs are geared towards less aggressive failure scenarios than LDCs they are considerably simpler and more efficient. We start be reviewing the motivation behind these codes.

5.1. Applications. Modern large scale distributed storage systems such as data centers store data in a redundant form to ensure reliability against node (e.g., individual machine) failures. The simplest solution here is the straightforward replication of data packets across different nodes. Alternative solution involves erasure coding: the data is partitioned into k information packets. Subsequently, using an erasure code, $N-k$ parity packets are generated and all N packets are stored in different nodes.

Using erasures codes instead of replication may lead to dramatic improvements both in terms of redundancy and reliability. However to realize these improvements one has to address the challenge of maintaining an erasure encoded representation. In particular, when a node storing some packet fails, one has to be able to quickly reconstruct the lost packet in order to keep the data readily available for the users and to maintain the same level of redundancy in the system. We say that a certain packet has *locality* r if it can be recovered from accessing only r other packets. One way to ensure fast reconstruction is to use erasure codes where all packets have low locality $r \ll k$. Having small value of locality is particularly important for information packets.

These considerations lead to introduction of (r, d)-local reconstruction codes, i.e., a linear codes capable of correcting any $d-1$ erasures where all information symbols have locality at most r. Storage systems based on (r, d)-codes provide fast recovery of information packets from a single node failure (typical scenario), and ensure that no data is lost even if up to $d-1$ nodes fail simultaneously.

5.2. Structure of LRCs. We begin by introducing some basic notions. A linear code is a linear mapping $C : \mathbb{F}_q^k \to \mathbb{F}_q^N$, where $k \leq N$. Every such mapping can be represented as

$$C(\mathbf{x}) = (\mathbf{x} \cdot \mathbf{p}_1, \ldots, \mathbf{x} \cdot \mathbf{p}_N), \tag{5.1}$$

where $\mathbf{p}_1, \ldots, \mathbf{p}_N \in \mathbb{F}_q^k$. We say that C is a systematic code if all for all $i \in [k]$, $\mathbf{p}_i$ is the i-th unit vector, i.e., the vector whose unique non-zero coordinate i carries value 1. In other words, a code is systematic if it performs encoding by appending redundant symbols to the original message. We refer to coordinates 1 through k of a systematic code as information coordinates.

We say that the i-th coordinate of C has locality r if, when erased, this coordinate can be recovered by accesing at most r of the $N-1$ remaining coordinates of a codeword. This is equivalent to saying that the vector $\mathbf{p}_i$ in (5.1) is in the span of some r vectors of

$\{\mathbf{p}_j\}_{j\in[N]\setminus\{i\}}$. Further we say that a systematic code C has information locality r, if all information coordinates of C have locality r. Finally we say that a code C has distance d, if C corrects any pattern of up to $d-1$ simultaneous erasures.

Definition 5.1. A linear systematic code $C : \mathbb{F}_q^k \to \mathbb{F}_q^N$ that has distance d and information locality r is called an (r, d)-local reconstruction code.

Below we present one simple family of (r, d)-local reconstruction codes, called Pyramid codes. We assume $r \mid k$.

Pyramid codes. To define an (r, d)-Pyramid code C encoding messages of dimension k we fix an arbitrary linear systematic code $E : \mathbb{F}_q^k \to \mathbb{F}_q^N$ that has distance d and codeword length $N = k - d + 1$. Note that such a code always exist provided that $q \geq N - 1$. Let

$$E(\mathbf{x}) = (\mathbf{x}, \mathbf{p}_0 \cdot \mathbf{x}, \mathbf{p}_1 \cdot \mathbf{x}, \ldots, \mathbf{p}_{d-2} \cdot \mathbf{x}).$$

We partition the set $[k]$ into $t = \frac{k}{r}$ disjoint subsets of size r, $[k] = \bigsqcup_{j\in[t]} S_j$. For a k-dimensional vector $\mathbf{x}$ and a set $S \subseteq [k]$ let $\mathbf{x}|_S$ denote the $|S|$-dimensional restriction of $\mathbf{x}$ to coordinates in the set S. We define the systematic code C by

$$C(\mathbf{x}) = (\mathbf{x}, (\mathbf{p}_0|_{S_1} \cdot \mathbf{x}|_{S_1}), \ldots, (\mathbf{p}_0|_{S_t} \cdot \mathbf{x}|_{S_t}), \ \mathbf{p}_1 \cdot \mathbf{x}, \ldots, \mathbf{p}_{d-2} \cdot \mathbf{x}.)$$

It is not hard to verify that the code C has distance d. We now argue that each information symbol $i \in [k]$ has locality r. Consider an arbitrary $i \in S_j$. Note that the value of $\mathbf{x}(i)$ can be deduced from accessing the light parity $\mathbf{p}_0|_{S_j} \cdot \mathbf{x}|_{S_j}$ and the values of information symbols $\mathbf{x}(l)$ for $l \in S_j \setminus \{i\}$.

Interestingly the simple construction above yields (r, d)-LRCs of the lowest possible redundancy.

Theorem 5.2. *For any linear code $C : \mathbb{F}_q^k \to \mathbb{F}_q^N$ of distance d and information locality r,*

$$N \geq k + \left\lceil \frac{k}{r} \right\rceil + d - 2. \tag{5.2}$$

5.3. Maximal recoverability. A stronger version of Theorem 5.2 shows that under some minor technical assumptions all (r, d)-LRCs of the lowest possible redundancy are in a certain sense very similar to Pyramid codes. In particular, such codes have the same *topology*, i.e., the same set of dependency relations between information symbols and parity symbols. Specifically, assuming $r \mid k$:

- Data symbols are partitioned into k/r groups of size r. For each such group there is one (local) parity symbol that stores the XOR (or some other non-trivial linear combination) of respective data symbols.
- The remaining $h = d - 2$ (heavy) parity symbols depend on all k data symbols.

In what follows we refer to codes meeting the description above as data-local (k, r, h)-codes. We also refer to a group of r data symbols and their local parity as a local group. (r, d)-LRCs with optimal redundancy are instances of data-local (k, r, h)-codes with $h = d - 2$.

Note that the class data-local codes is fairly broad as there is a lot of flexibility is choosing coefficients in heavy parities. All these codes have appropriate information locality. However

they differ in terms of reliability guarantees that they provide as correctability of a particular failure pattern obviously depends on coefficients used to define heavy parities.

We say that a data-local (k, r, h)-code is Maximally Recoverable (MR) if it corrects every failure patterns that is correctable by some other code that has the same topology. Another equivalent way to define maximally recoverability is as follows:

Definition 5.3. Let C be a data-local (k, r, h)-code. We say that C is maximally recoverable if it corrects any failure pattern that can be obtained by erasing one coordinate in each of $\frac{k}{r}$ local groups as well as h arbitrary additional coordinates.

Note that maximal recoverability is a much stronger property than the mere distance required in the definition of (r, d)-local reconstruction codes.

It is not hard to show that maximally recoverable codes exist. In fact, simply picking coefficients of heavy parities at random from a large enough finite field with high probability yields an MR code. However in applications we would like to have codes over small finite fields to facilitate fast encoding and decoding. The best explicit constructions of such codes are given by the following

Theorem 5.4. *For constants r and h and for all k such that $r \mid k$, there exists a maximally recoverable data-local (k, r, h)-code over a field of size $O\left(k^{\lceil (h-1)\left(1-\frac{1}{2^r}\right)\rceil}\right)$.*

5.4. Notes. Pyramid codes were introduced in [16]. General local reconstruction codes were studied in [10]. Theorem 5.4 is from [11]. See also [3]. Local reconstruction codes are used in practice. Instances of these codes were first deployed by Windows Azure Storage [17], and have later been used in a number of other production systems. A different other notion of local reconstruction in codes for storage has been addressed in [8].

The main open challenge in the area of local reconstruction codes is to reduce the field size of maximally recoverable codes. The best upper bound for the field size is roughly $O(k^{h-1})$ while the only available lower bound is $\Omega(k)$ independent of h. Constructing explicit maximally recoverable codes over small finite fields in more general topologies is also of great interest.

6. Conclusion

In this survey we reviewed two main families of codes with local decoding procedures, namely locally decodable codes and local reconstructions codes. There is a large array of questions that remain open. In the case of LDCs the main open questions pertain to the true shape of the tradeoff between codeword length and query complexity. In the case of LRCs this tradeoff is understood and the main challenges are in constructing maximally recoverable codes over small finite fields. There is also a large area dealing with codes that provide local recovery of message symbols after more than one but less than $\Omega(N)$ erasures and thus bridge LDCs and LRCs. While there are some well studied families of codes that fall in this range, e.g., projective geometry codes [22], in general this regime is not well understood.

References

[1] Barrington, D. A., Beigel, R., and Rudich, S., *Representing Boolean functions as polynomials modulo composite numbers*, Computational Complexity **4** (1994), 367–382.

[2] Ben-Aroya, A., Efremenko, K., and Ta-Shma, A., *Local list decoding with a constant number of queries*, In Proceedings of the 51st IEEE Symposium on Foundations of Computer Science (FOCS) (2010), 715–722.

[3] Blaum, M., Hafner, J. L., and Hetzler, S., *Partial-MDS codes and their application to RAID type of architectures*, IEEE Transactions on Information Theory **59** (7) (2013), 4510-4519.

[4] Chor, B., Goldreich, O., Kushilevitz., E., and Sudan, M., *Private information retrieval*, Journal of the ACM **45** (1998), 965–981.

[5] Dvir, Z., *Incidence theorems and their applications*, Foundations and trends in theoretical computer science **6** (4) (2012), 257–393.

[6] Dvir, Z., Gopalan, P., and Yekhanin, S., *Matching vector codes*, SIAM Journal on Computing **40** (4) (2011), 1154–1178.

[7] Dvir, Z., Kopparty, S., Saraf., S., and Sudan, M., *Extensions to the method of multiplicities, with applications to Kakeya sets and mergers*, SIAM Journal on Computing **42** (6) (2013), 2305–2328.

[8] Dimakis, A. G., Godfrey, B., Wu, Y., Wainwright, M. J., and Ramchandran, K., *Network coding for distributed storage systems*, IEEE Transactions on Information Theory **56** (2010), 4539–4551.

[9] Efremenko, K., 3*-query locally decodable codes of subexponential length*, SIAM Journal on Computing **41** (6) (2012), 1694–1703.

[10] Gopalan, P., Huang, C., Simitci, H., and Yekhanin, S., *On the locality of codeword symbols*, IEEE Transactions on Information Theory **58** (11) (2012), 6925–6934.

[11] Gopalan, P., Huang, C., Jenkins, B., and Yekhanin, S., *Explicit maximally recoverable codes with locality*, Proceedings of the Electronic Colloquium on Computational Complexity (ECCC) **20** (2013).

[12] Grolmusz, V., *Superpolynomial size set-systems with restricted intersections mod* 6 *and explicit Ramsey graphs*, Combinatorica **20** (2000), 71–86.

[13] Guo, A., Kopparty, S., and Sudan, M., *New affine-invariant codes from lifting.*, In Proceedings of the 4th Conference on Innovations in Theoretical Computer Science (ITCS) (2013), 529–540.

[14] Hamming, R. W., *Error detecting and error correcting codes*, The Bell System Technical Journal **26** (2) (1950), pp. 147–160.

[15] Hemenway, B., Ostrovsky, R., and Wootters, M., *Local correctibility of expander codes*, In Proceedings of the 40th International Colloquium on Automata, Languages, and Programming (ICALP) (2013), 540–551.

[16] Huang, C., Chen, M., and Li, J., *Pyramid Codes: flexible schemes to trade space for access efficiency in reliable data storage systems*, In Proceedings of 6th IEEE International Symposium on Network Computing and Applications (NCA) (2007), 79–86.

[17] Huang, C., Simitci, H., Xu, Y., Ogus, A., Calder, B., Gopalan, P., Li, J., and Yekhanin, S., *Erasure coding in Windows Azure Storage*, In Proceedings of the USENIX Annual Technical Conference (USENIX ATC) (2012), 15–27.

[18] Katz, J. and Trevisan, L., *On the efficiency of local decoding procedures for error-correcting codes*, In Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC) (2000), 80–86.

[19] Kerenidis, I., and de Wolf, R., *Exponential lower bound for* 2*-query locally decodable codes via a quantum argument*, Journal of Computer and System Sciences **69** (2004), 395–420.

[20] Kopparty, S., Saraf, S., and Yekhanin, S., *High-rate codes with sublinear-time decoding*, In Proceedings of the 43nd ACM Symposium on Theory of Computing (STOC) (2011), 176–176.

[21] Muller, D. E., *Application of boolean algebra to switching circuit design and to error detection*, IEEE Transactions on Computers **3** (1954), 6–12.

[22] Peterson, W. W., and Weldon, E. J., *Error correcting codes*. Princeton Mathematical Series 39, The MIT Press, Cambridge, MA, 1972.

[23] Reed, I. S., *A class of multiple-error-correcting codes and the decoding scheme*, IEEE Transactions on Information Theory **4** (1954), 38–49.

[24] Shannon, C. E., *A mathematical theory of communication*, The Bell System Technical Journal **27** (1948), pp. 379–423, 623–656.

[25] Sudan, M., Trevisan, L., and Vadhan, S., *Pseudorandom generators without the XOR lemma*, In Proceedings of the 31st ACM Symposium on Theory of Computing (STOC) (1999), 537–546.

[26] Woodruff, D., *A quadratic lower bound for three query linear locally decodable codes over any field*, In Proceedings of the International Workshop on Randomization and Computation (RANDOM) (2010), 766–779.

[27] Yekhanin, S., *Private information retrieval*, Communications of the ACM **53** (4) (2010), 68–73.

[28] ———, *Towards* 3*-query locally decodable codes of subexponential length*, Journal of the ACM **55** (2008), 1–16.

[29] ———, *Locally decodable codes*, Foundations and trends in theoretical computer science **6** (3) (2012), 139–255.

Microsoft Research, 1065 La Avenida, Mountain View, 94043, USA.
E-mail: yekhanin@microsoft.com

15. Numerical Analysis and Scientific Computing

On a class of high order schemes for hyperbolic problems

Rémi Abgrall

Abstract. This paper provides a review about a family of non oscillatory and parameter free finite element type methods for advection-diffusion problems. Due to space limitation, only the scalar hyperbolic problem is considered. We also show that this class of schemes can be interpreted as finite volume schemes with multidimensional fluxes.

Mathematics Subject Classification (2010). 65, 76.

Keywords. Numerical approximation of hyperbolic problems, Non oscillatory schemes, Unstructured meshes, High order methods

1. Introduction

We are interested in the numerical solution of parabolic type equations in which the elliptic terms play an important role only at some locations of the computational domain. To make things more precise, our target are the Navier-Stokes equations in the compressible regime. These systems of partial differential equations are supplemented by initial and boundary conditions. In particular, at solid walls, the velocity is set to zero and the temperature behavior is specified. Thus depending on the Reynolds number, the viscous terms have an effect that is sensitive on a more or large range. Far enough from the walls, where the viscous effects are less prominent, it is mainly the hyperbolic part that plays the major role, and thus, depending on the flow conditions, thin zone with very steep gradients may exist with a shock-like structure.

Our goal is to approximate the solution every where, with a parameter free method, so that the solution is oscillation free, with a uniform accuracy. In addition, we want to handle complicated geometries, so that the method use unstructured meshes.

How can this program be achieved? In the following, we focus on steady problems, and to make things simpler, we focus on the scalar problem:

$$\text{div } \mathbf{f}(u) = 0 \tag{1a}$$

subjected to

$$\min(\nabla_u f(u) \cdot \mathbf{n}(\mathbf{x}), 0)(u - g) = 0 \text{ on } \partial\Omega \tag{1b}$$

In (1b), $\mathbf{n}(\mathbf{x})$ is the outward unit vector at $\mathbf{x} \in \partial\Omega$ (thus we assume enough regularity for Ω). The case of the advection-diffusion problem

$$\text{div } \mathbf{f}(u) - \text{div}(\mathbf{K}\nabla u) = 0 \tag{2a}$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

subjected to boundary condition of the Dirichlet type

$$u = g \text{ on } \Gamma_D \tag{2b}$$

and Neuman-like conditions

$$(\mathbf{K}\nabla u)\cdot \mathbf{n}(\mathbf{x}) = h(\mathbf{x}) \text{ on } \Gamma_N \tag{2c}$$

is done in a similar way except for some technicalities about the diffusion term, see [4]. Extensions to the system case can be found in [5] for the pure hyperbolic case and [3] for the Navier Stokes equations.

Here the notations are standard: g and h are regular enough functions, Γ_D and Γ_N are non overlapping regular enough subsets of $\partial\Omega$, and $\Gamma_D \cup \Gamma_N = \partial\Omega$. From now on, we assume that Ω has a polyhedric boundary, and more over $\Omega_h = \Omega$ for the chosen family of triangulations in order to simplify. These assumptions are by no mean essential. We denote by $\mathcal{E}_h$ the set of edges/faces of $\mathcal{T}_h$ that are contained in $\partial\Omega$, and $\mathcal{K}$ stands either for an element K or a face/edge e.

In the finite element setting, there exists several variational formulations of this class of problems. The classical ones can be defined in three steps. We are given a family of meshes denoted by $(\mathcal{T}_h)_{h\in\mathcal{H}}$. These meshes are made of elements denoted generically by K. The parameter h, as usual, denotes the maximum of the diameters of K, $K \in \mathcal{T}_h$. The meshes can be geometrically conformal or not. Then we need to define the trial function space, denoted by U_h and a test function V_h. The last step is to define a bi-linear form a on $U_h \times V_h$, as well as form ℓ defined on V_h. As usual, we assume that the spaces U_h and V_h encode some of the boundary conditions, while the others are encoded in ℓ. The problem is to find $u_h \in U_h$ such that a for any $v_h \in V_h$, we have

$$a(u_h, v_h) = \ell(v_h).$$

A first example example is given by the streamline diffusion method [12, 13] for which there are two possible interpretations. In the first one, we consider a Petrov Galerkin formulation, .i.e we take

$$U_h = \{u_h \in H^1(\Omega) \text{ such that for any } K \in \mathcal{T}_h, \quad u_h\big|_K \in \mathbb{P}^r(K)\} \cap C^0(\overline{\Omega})$$

and

$$\begin{aligned} V_h = &\{v_h \in L^2(\Omega), \text{ such that for any } K \in \mathcal{T}_h, \text{ there exists } w_h \in U_h, \\ &v_h = w_h + h_K \tau_K \nabla_u \mathbf{f}(u_h)\nabla w_h\} \end{aligned}$$

and

$$a_{\text{SUPG1}}(u_h, v_h) = \int_\Omega v_h \ \text{div}\, \mathbf{f}(u_h) + \sum_{e\in\mathcal{E}_h} \int_e v_h \big(\hat{\mathbf{f}}_{\mathbf{n}}(g, u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}\big) = \int_\Omega f v_h. \tag{3a}$$

Here, $\hat{\mathbf{f}}$ is a consistent upwind numerical flux. The second interpretation is to take $V_h = U_h$ and use, instead of a_{SUPG1} the form a_{SUP2} defined by

$$a_{\text{SUPG2}}(u_h, v_h) = \int_\Omega v_h \ \text{div}\, \mathbf{f}(u_h) + \sum_K h_K \int_K \big(\nabla_u f(u_h)\nabla v_h\big)\tau_K\big(\nabla_u f(u_h)\nabla u_h\big)$$

$$+ \sum_{e\in\mathcal{E}_h} \int_e v_h \big(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}\big). \tag{3b}$$

This can be as a Galerkin approximation of a modified equation, namely

$$\text{div } \mathbf{f}(u) - \text{div } \Big(h\nabla_u \mathbf{f}(u)\otimes(\tau\nabla_u \mathbf{f}(u))\nabla u\Big) = 0. \tag{3c}$$

In (3), the parameters τ_K are positive functions (typically constant per element) and in (3c) the function τ and h are defined by their restrictions on each element.

We can play further with the trial and test spaces. If one removes the continuity assumption, then we have a Discontinuous Galerkin formulation, i.e. $U_h = V_h$ with

$$U_h = \{u_h \in L^2(\Omega),\ \text{such that for any } K\in\mathcal{T}_h, \quad u_h\big|_K \in \mathbb{P}^r(K)\}$$

and

$$a(u_h,v_h) = \sum_{K\in\mathcal{T}_h} \Big(-\int_K \nabla v_h\cdot\mathbf{f}(u^h) + \int_{\partial K} v_h \hat{\mathbf{f}}_{\mathbf{n}}\big((u_h)_{|K},(u_h)_{|K^-}\big)\Big) \tag{4a}$$

where K^- denotes generically the element(s) that are on the other side of the faces of ∂K. Another formulation is

$$\begin{aligned} a(u_h,v_h) = &\sum_{K\in\mathcal{T}_h} \Big(-\int_K \nabla v_h\cdot\mathbf{f}(u^h) + \int_{\partial K} v_h \hat{\mathbf{f}}_{\mathbf{n}}\big((u_h)_{|K},(u_h)_{|K^-}\big)\Big) \\ &+ \sum_K h_K \int_K \big(\nabla_u f(u_h)\nabla v_h\big)\tau_K\big(\nabla_u f(u_h)\nabla u_h\big) \end{aligned} \tag{4b}$$

In (4), the Dirichlet boundary conditions are set weakly, as in (3), by setting $u_h = g$ on the parts of ∂K which belongs to inflow part of $\partial\Omega$.

The space U_h and V_h can be independently chosen, as well as a and ℓ, provided the variational problem is consistent with the problem (1), and of course the numerical method is stable. Formal accuracy is obtained via the choice the polynomial degree r, and effective accuracy is related to the stability of the scheme in suitable norm. Hence a natural question is: can we define U_h, V_h and the forms a and ℓ such that in addition with consistency and accuracy, we can also have non oscillatory properties. In the case of the streamline methods, this last property is obtained by modifying the formulation by adding a dissipation operator which is parameter dependent. In the case of the Discontinuous Galerkin method, this property is obtained via a proper choice of the arguments in $\hat{\mathbf{f}}_{\mathbf{n}}$, see [7, 8]. We note that only the averages in K are controlled. In both cases this is obtained by introducing some genuine non linearity in the scheme, i.e. even if (1) is a linear problem, the scheme will be non linear.

In this paper, we show that, by introducing a solution-dependent operator χ from $U_h\cap C^0(\Omega)$ to $L^2(\Omega)$, the variational problem with a defined by

$$a(u_h,v_h) = \sum_K \int_K \chi_u^h(v_h)\text{div } \mathbf{f}(u^h) + \sum_{e\in\mathcal{E}} \int_e v_h(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}) \tag{5}$$

enables to get all the properties. The rest of this paper is organized as follow: inspired by a rewriting of (3), we introduce the residual distribution schemes. We provide a simple criteria which guaranty a Lax-Wendroff type theorem, provide a simple criteria that guaranties

formal accuracy, show how the choice of norms guaranty the effective accuracy, and provide several examples of schemes. In the last part, we show how these schemes can also be interpreted as finite volume schemes and we provide explicit formula.

2. Formulation of residual distribution schemes

These schemes have been introduced by P.L. Roe in [17] in one dimension, and [18] in the multidimensional case. As we see, there are many common points with the streamline method, the difference is that we try to combine ideas from the finite element community and from the finite volume one. The first scheme of this kind was probably designed by R. Ni [16] where introduce a particular version of the Lax Wendroff scheme.

2.1. Definition, connection to finite element methods. We make the standard remark that, for any internal degree of freedom σ, if φ_σ is the Lagrange basis function associated to σ, (3b) can be written as:

$$a_{\text{SUPG2}}(u_h,\varphi_\sigma)=\sum_K\Big(\int_K\varphi_\sigma\nabla\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\Big)$$
$$+\sum_{e\in\mathcal{E}_h}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}).$$

Since the support of φ_σ is made of all the elements K that share σ, we have for any degree of freedom σ:

$$a_{SUPG2}(u_h,\varphi_\sigma)$$
$$=\sum_{K\ni\sigma}\Big(\int_K\varphi_\sigma\nabla\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\Big)$$
$$\sum_{e\in\mathcal{E}_h,\sigma\in e}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n})$$

and notice that

1. for any K,

$$\sum_{\sigma\in K}\Big(\int_K\varphi_\sigma\nabla\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\Big)=\int_{\partial K}\mathbf{f}(u_h)\cdot\mathbf{n},$$

2. for any $e\in\mathcal{E}_h$,

$$\sum_{\sigma\in e}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n})=\int_e(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}).$$

This is true because $\sum_{\sigma\in K}\varphi_\sigma(\mathbf{x})=1$ and thus $\sum_{\sigma\in K}\nabla\varphi_\sigma(\mathbf{x})=0$ for all $x\in\mathcal{K}$.

We define the total residual for element and edges the quantities:

$$\Phi^K:=\int_{\partial K}\mathbf{f}(u_h)\cdot\mathbf{n},\ \text{and}\ \Phi^e:=\int_e(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}).\tag{6}$$

A residual distribution scheme is defined by the sub-residuals that are "sent" to the degrees of freedom σ by an element K (resp. a boundary edge e). We denote them by $\Phi_\sigma^K(u^h_{|K})$ (resp. $\Phi_\sigma^e(u^h_{|e})$). The scheme writes, for any internal degree of freedom σ,

$$\sum_{K\ni\sigma} \Phi_\sigma^K(u^h_{|K}) = 0, \tag{7a}$$

and for any degree of freedom on the boundary,

$$\sum_{K\ni\sigma} \Phi_\sigma^K(u^h_{|K}) + \sum_{e\ni\sigma} \Phi_\sigma^e(u^h_{|e}) = 0. \tag{7b}$$

We assume that the following structure condition holds true:

$$\forall \sigma \in K, \qquad \sum_{\sigma\in K} \Phi_\sigma^K(u^h_{|K}) = \int_{\partial K} \mathbf{f}(u_h)\cdot\mathbf{n} \qquad (=\Phi^K), \tag{8a}$$

$$\forall e \in \mathcal{E}_h, \qquad \sum_{\sigma\in e} \Phi_\sigma^e(u^h_{|K}) = \int_e (\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}). \qquad (=\Phi^e) \tag{8b}$$

We see that the SUPG method is a particular case of such scheme. There is a lot of freedom in defining the sub-residuals $\Phi_\sigma^K(u^h_{|K})$ and $\Phi_\sigma^e(u^h_e)$, we will show how we can take advantage of this freedom to achieve our goal. Note that in the definition of the sub-residual, we have implicitly assumed that only the degrees of freedom with K or e are necessary to define these quantities: the stencil of the method is the most possible compact which is a good point for the parallelization of the method.

Another example of sub-residual are the Galerkin residuals defined by: on the element K

$$\Phi_\sigma^{G,K} = \int_K \varphi_\sigma \text{div } \mathbf{f}(u^h = -\int_K \nabla\varphi_\sigma\cdot\mathbf{f}(u^h) + \int_{\partial K} \varphi_\sigma \mathbf{f}(u^h)\cdot\mathbf{n}, \tag{9a}$$

and on the boundary face e:

$$\Phi_\sigma^{G,e} = \int_e \varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}) \tag{9b}$$

We see that both $\{\Phi_\sigma^{G,K}\}_{\sigma\in K}$ and $\{\Phi_\sigma^{G,e}\}_{\sigma\in e}$ satisfy (8) with the same value of the total residual. Unfortunately, the scheme (7) with the Galerkin residual is widely unstable,

2.2. Structure conditions. For any w^h (not necessarily a solution of (7) if it exists), and any test function v^h, we have (setting $v^h_\sigma = v^h(\sigma)$):

$$\begin{aligned} S &:= \sum_{\sigma\notin\partial\Omega} v^h_\sigma \left(\sum_{K\ni\sigma} \Phi_\sigma^K(w^h_{|K}) \right) + \sum_{\sigma\in\partial\Omega} v^h_\sigma \left(\sum_{K\ni\sigma} \Phi_\sigma^K(w^h_{|K}) + \sum_{e\ni\sigma, e\in\mathcal{E}_h} \Phi_\sigma^e(w^h_{|e}) \right) \\ &= \sum_K \left(\sum_{\sigma\in K} v^h_\sigma \Phi_\sigma^K(w^h_{|K}) \right) + \sum_{e\in\mathcal{E}_h} \left(\sum_{\sigma\in e} v^h_\sigma \Phi_\sigma^e(w^h_{|K}) \right) \\ &= -\int_\Omega v^h \nabla\cdot\mathbf{f}(u^h) + \int_{\partial\Omega} v^h \hat{\mathbf{f}}_{\mathbf{n}}(g,w^h) \end{aligned} \tag{10}$$

$$+\sum_K \sum_{\sigma\in K} v_\sigma^h\big(\Phi_\sigma^K(w^h_{|K}) - \Phi_\sigma^{G,K}(w^h_{|K})\big)$$
$$+\sum_{e,e\in\mathcal{E}_h}\sum_{\sigma\in e} v_\sigma^h\big(\Phi_\sigma^e(w^h_{|K}) - \Phi_\sigma^{G,e}(w^h_{|K})\big)$$

thanks to (9). Then, since (recall $\mathcal{K}$ represents either a generic element or a generic member of $\mathcal{E}_h$)

$$\sum_{\sigma\in\mathcal{K}}\left(\Phi_\sigma^{\mathcal{K}}(w^h_{|\mathcal{K}}) - \Phi_\sigma^{G,\mathcal{K}}(w^h_{|K})\right) = 0,$$

(10) becomes, denoting by n_K and n_e the number of degree of freedom in K and e:

$$\begin{aligned} S = &-\int_\Omega \nabla v^h\cdot\mathbf{f}(u^h) + \int_\Omega v^h\hat{\mathbf{f}}_{\mathbf{n}}(g,w^h)\\ &+\sum_K \frac{1}{n_K!}\sum_{\sigma,\sigma'\in K}\big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^K(w^h_{|K}) - \Phi_\sigma^{G,K}(w^h_{|K})\big)\\ &+\sum_{e\in\mathcal{E}_h}\frac{1}{n_e!}\sum_{\sigma,\sigma'\in e}\big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^e(w^h_{|e}) - \Phi_\sigma^{G,e}(w^h_{|e})\big)\end{aligned} \tag{11}$$

This relation is fundamental in our analysis.

2.2.1. Conservation. In [6], we prove the following result:

Theorem 2.1. *Assume the family of meshes $\mathcal{T} = (\mathcal{T}_h)_{h\in\mathcal{H}}$ is regular. We assume that the residuals $\{\Phi_\sigma^{\mathcal{K}}\}_{\sigma\in\mathcal{K}}$, for $\mathcal{K}$ an element or a boundary element of $\mathcal{T}_h$, satisfy:*

- *For any $M\in\mathbb{R}^+$, there exists a constant C which depends only on the family of meshes $\mathcal{T}_h$ and M such that for any $u_h\in U_h$ with $||u^h||_\infty\le M$, then*

$$\left\|\Phi_\sigma^{\mathcal{K}}(u^h{}_{|\mathcal{K}})\right\| \le C\sum_{\sigma,\sigma'\in\mathcal{K}}|u_\sigma^h - u_{\sigma'}^h|$$

- *they satisfy the conservation property* (8)*.*

Then if there exists a constant C_{max} such that the solutions of the scheme (7) *satisfy $||u^h||_\infty\le C_{max}$ and a function $v\in L^2(\Omega)$ such that $(u^h)_h$ or at least a sub-sequence converges to v in $L^2(\Omega)$, then v is a weak solution of* (1)

Proof. The proof can be found in [6], it uses (11) and some adaptation of the ideas of [14]. □

We can also state condition for entropy inequalities:

Proposition 2.2. *Let $(U,\mathbf{G})$ be an couple entropy-flux for* (1) *and $\hat{\mathbf{G}}_{\mathbf{n}}$ an upwind numerical entropy flux consistent with $\mathbf{G}\cdot\mathbf{n}$. Assume that the residuals satisfy: for any element K,*

$$\sum_{\sigma\in K} U(u_\sigma)\Phi_\sigma^K \le \int_{\partial K}\mathbf{G}(u^h_{|\mathcal{K}})\cdot\mathbf{n} \tag{12a}$$

and for any boundary edge e,

$$\sum_{\sigma\in e} U(u_\sigma)\Phi^e_\sigma \le \int_e \big(\hat{\mathbf{G}}_{\mathbf{n}}(u^h_{|e}, g) - \mathbf{G}(u^h_{|K})\cdot\mathbf{n}\big). \tag{12b}$$

Then, under the assumptions of the theorem 2.1, the limit weak solution also satisfies the following entropy inequality: for any $\varphi \in C^1(\Omega)$, $\varphi \ge 0$,

$$-\int_\Omega \nabla\varphi\cdot\mathbf{G}(u) + \int_{\partial\Omega} \hat{\mathbf{G}}_{\mathbf{n}}(u, g) \le 0.$$

Proof. The proof is similar to that of theorem 2.1. □

2.2.2. Accuracy. In most cases, assuming a smooth solution of (1), the formal accuracy analysis is done by checking how large is the error made when plugging the exact solution into the scheme. This is carried out using Taylor expansions, and the geometry of the computational stencil plays an important role. When the mesh has no particular symmetry, this leads to nowhere. Instead of looking to how far the numerical scheme departs from the strong form of the PDE, it is much more flexible to look at how for it departs its weak form, i.e. instead of checking $\text{div}\,\mathbf{f}(u) = 0$, it is better to test, for any φ smooth enough, $\int_\Omega \varphi\,\text{div}\,\mathbf{f}(u) = 0$, of course after using the Green formula.

In practice, we define the truncation error

$$\mathcal{E}(w^h, v^h) = \sum_{\sigma\notin\partial\Omega} v^h_\sigma \left(\sum_{K\ni\sigma} \Phi^K_\sigma(w^h_{|K})\right),$$

and consider

$$\mathcal{E}(w^h) = \max_{v^h\in V^h, ||v^h||_{W^{1,\infty}}=1} \mathcal{E}(w^h, v^h). \tag{13}$$

We can then extend the classical definition of accuracy:

Definition 2.3 (Accuracy). We say that the scheme (7) is $r+1$-th order accurate if, for any smooth solution $u_{ex} \in C^{r+1}(\overline{\Omega})$ of (1), $\mathcal{E}(u^h_{ex}) \le C\ h^{r+1}$. The constant C only depend on the family $\mathcal{T}$, the regularity of $\mathbf{f}$, on the $r+1$ derivative of u, and the boundary conditions.

Using (11), we see that, for any v^h

$$\mathcal{E}(u^h_{ex}, v^h) = -\int_\Omega \nabla v^h\cdot\mathbf{f}(u^h_{ex}) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u^h_{ex}) \tag{14}$$

$$+\sum_K \frac{1}{n_K!} \sum_{\sigma,\sigma'\in K} \big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi^K_\sigma((u^h_{ex})_{|K}) - \Phi^{G,K}_\sigma((u^h_{ex})_{|K})\big) \tag{15}$$

$$+\sum_{e\in\mathcal{E}_h} \frac{1}{n_e!} \sum_{\sigma,\sigma'\in e} \big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi^e_\sigma((u^h_{ex})_{|K}) - \Phi^{G,e}_\sigma((u^h_{ex})_{|e})\big) \tag{16}$$

For the *steady* problem (1), we have the following result:

Lemma 2.4. *Let us recall that $\Omega \subset \mathbb{R}^d$ and is bounded.*

If the solution u_{ex} of the steady problem (1) *is C^{r+1}, then*

(1) $\Phi_\sigma^{G,K}((u_{ex}^h)_{|K}) = O(h^{r+d})$,

(2) $\Phi_\sigma^{G,e}((u_{ex}^h)_{|e}) = O(h^{r+d-1})$

(3) *if the numerical flux* $\hat{\mathbf{f}}$ *is Lipschitz,* $-\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}^h) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u_{ex}^h) = O(h^{r+1})$,

Proof. We start by showing the first result. The proof of the second one is similar and is omitted.

Since $u_{ex} \in C^{r+1}$, we have div $\mathbf{f}(u_{ex}) = 0$ in a strong sense, thus for any $K \in \mathcal{T}_h$ and any σ,

$$\int_K \varphi_\sigma \text{ div } \mathbf{f}(u_{ex}) = -\int_K \nabla\phi_\sigma \cdot \mathbf{f}(u_{ex}) + \int_{\partial K} \phi_\sigma \mathbf{f}(u_{ex}) \cdot \mathbf{n} = 0.$$

We can subtract this relation to $\Phi_\sigma^{G,K}(u_{ex}^h)$ and get:

$$\Phi_\sigma^{G,K}(u_{ex}^h) = -\int_K \nabla\varphi_\sigma \cdot \Big(\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_e)\Big) + \int_{\partial K} \varphi_\sigma \Big(\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_e)\Big).$$

Since the mesh is regular, we have:

$$|K| = O(h^d), \qquad \nabla\varphi_\sigma = O(h^{-1}), \qquad |\partial K| = O(h^{d-1})$$

and since the flux $\mathbf{f}$ is C^1, we have

$$\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_e) = O(h^{k+1}).$$

Gathering the pieces together, we get:

$$\Big|\Phi_\sigma^{G,K}(u_{ex}^h)\Big| \le C\Big(h^d \times h^{-1} \times h^{k+1} + h^{d-1} \times 1 \times h^{k+1}\Big) = O(h^{k+d}).$$

The third inequality is obtained in a similar manner: From (1), we have for any v^h, setting $\Gamma^- = \{\mathbf{x} \in \partial\Omega, \nabla_u \mathbf{f}(u) \cdot \mathbf{n} < 0\}$,

$$-\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}) + \int_{\Gamma^-} v^h \mathbf{f}(u_{ex}) \cdot \mathbf{n} = 0$$

so that

$$\begin{aligned} -\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}^h) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u_{ex}^h) & \\ = -\int_\Omega \nabla v^h \cdot \big(\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_{ex})\big) + \int_{\partial\Omega} v^h \Big(\hat{\mathbf{f}}_{\mathbf{n}}(g, u_{ex}^h) - \mathbf{f}(u_{ex}^h) \cdot \mathbf{n}\Big) & \\ = (I) + (II) & \end{aligned}$$

Using again the same arguments, since the numerical flux is Lipschitz continuous, we see that both (I) and (II) are of the order of $O(h^{k+1}) \times ||v^h||_{W^{1,\infty}(\Omega)}$. □

Then, we have:

Proposition 2.5. *Under the assumptions of Lemma 2.4 and assuming that the family of meshes* $\mathcal{T}_h$ *is regular, the residuals satisfy:*

$$\text{for all } \sigma \text{ and all } \mathcal{K} = K \text{ or } e, \Phi_\sigma^{\mathcal{K}}((u_{ex})_{|\mathcal{K}}) = O(h^{r+D}) \tag{17}$$

where $D = d$ *for elements and* $D = d-1$ *for* $e \in \mathcal{E}$*, then the scheme is formally* $r+1$ *accurate.*

Proof. $\mathcal{E}(u_{ex}^h, v^h)$ is the sum of

$$-\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}^h) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u_{ex}^h)$$

which is $O(h^{r+1})$ by lemma 2.4 and

$$\sum_K \frac{1}{n_K!} \sum_{\sigma,\sigma' \in K} \big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^K(w_{|K}^h) - \Phi_\sigma^{G,K}(w_{|K}^h)\big) + \sum_{e \subset \Omega} \frac{1}{n_e!} \sum_{\sigma,\sigma' \in e} \big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^e(w_{|K}^h) - \Phi_\sigma^{G,e}(w_{|K}^h)\big)$$

Since the mesh is regular, the number of elements in the mesh is $O(h^{-d})$ and the number of boundary elements is $O(h^{d-1})$. Since $v \in W^{1,\infty}$, its Lagrange interpolant satisfy

$$\big|v_\sigma^h - v_{\sigma'}^h\big| \leq h||v^h||_{W^{1,\infty}}$$

and $\sup_h ||v^h||_{W^{1,\infty}}$ is bounded by a constant that depends on $\mathcal{T}$ and $||v||_{1,\infty}$. Then we see that

$$\begin{aligned}
&\Big|\sum_K \frac{1}{n_K!} \sum_{\sigma,\sigma' \in K} \big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^K(w_{|K}^h) - \Phi_\sigma^{G,K}(w_{|K}^h)\big) \\
&\qquad + \sum_{e \subset \partial\Omega} \frac{1}{N_e!} \sum_{\sigma,\sigma' \in e} \big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^e(w_{|K}^h) - \Phi_\sigma^{G,e}(w_{|K}^h)\big)\Big| \\
&\leq C\big(h^{-d} \times h \times h^{d+r} + h^{-d+1} \times h \times h^{r+d-1}\big) \\
&\leq Ch^{r+1}
\end{aligned}$$

□

3. Construction of monotonicity preserving arbitrary accurate schemes

We start by a basic remark that goes at least back to A. Harten [11], and we rephrase it in the Residual Distribution framework.

Lemma 3.1. *Assume that the residual (for element and edges) write, for any degree of freedom,*

$$\Phi_\sigma^{\mathcal{K}}(u_h) = \sum_{\sigma' \ni \mathcal{K}} c_{\sigma\sigma'}^K (u_\sigma - u_{\sigma'}), \tag{18}$$

then the iterative scheme

$$u_\sigma^{n+1} = u_\sigma^n - \omega_\sigma \Big(\sum_{K \ni \sigma} \Phi_\sigma^K + \sum_{e \ni \sigma} \Phi_\sigma^e \big)$$

admits a local maximum principle if

- *for any* σ, σ', $c_{\sigma\sigma'}^K \geq 0$,
- $\omega_\sigma \left(\sum_{K \ni \sigma} \sum_{\sigma' \in K} c_{\sigma\sigma'}^K + \sum_{\sigma' \in K} c_{\sigma\sigma'} \right) \leq 1$

Proof. It is clear that:

$$\sum_{K\ni\sigma}\Phi_\sigma^K+\sum_{e\ni\sigma,e\in\mathcal{E}_h}\Phi_\sigma^e=\Big(\sum_{K\ni\sigma}\sum_{\sigma'\in K}c_{\sigma\sigma'}^K+\sum_{\sigma'\in K}c_{\sigma\sigma'}^K\Big)u_\sigma$$
$$+\sum_{\sigma'}\Big(\sum_{K,\sigma,\sigma'\in K}c_{\sigma\sigma'}^K\Big)u_{\sigma'}$$

Here, in order to simplify the notations, we have set $c_{\sigma,\sigma'}^{\mathcal{K}}=0$ when $\sigma\notin\mathcal{K}$ or $\sigma'\notin\mathcal{K}$.

The results holds true because $c_{\sigma\sigma'}^{\mathcal{K}}\geq 0$, and

$$\sum_{K\ni\sigma}\sum_{\sigma'\in\mathcal{K}}c_{\sigma\sigma'}^{\mathcal{K}}+\sum_{\sigma'\in\mathcal{K}}c_{\sigma\sigma'}^{\mathcal{K}}=\sum_{\sigma'}\Big(\sum_{\mathcal{K},\sigma,\sigma'\in\mathcal{K}}c_{\sigma\sigma'}^{\mathcal{K}}\Big).$$ □

The idea is to construct schemes that satisfy the requirement $c_{\sigma,\sigma'}^{\mathcal{K}}\geq 0$. It is known since Godunov that one cannot have a scheme that is both monotonicity preserving and high order accurate, hence some sort of non linearity must be introduced. Before showing how we can meet the requirements, let us introduce our reference monotone scheme. It is a multidimensional extension of the Rusanov (or local Lax-Friedrichs) scheme, namely, for any $\mathcal{K}$ and σ,

$$\Phi_\sigma^{\mathcal{K}}=\frac{1}{n_{\mathcal{K}}}\Phi^{\mathcal{K}}+\alpha_k\big(u_\sigma-\overline{u}_{\mathcal{K}}\big),\qquad \overline{u}_{\mathcal{K}}=\frac{1}{n_{\mathcal{K}}}\sum_{\sigma\in\mathcal{K}}u_\sigma \tag{19}$$

This scheme has the form (18) and is monotone if $\alpha_K\geq\max_{\mathcal{K}}||\nabla_u\mathbf{f}(u^h)||$.

Another example of monotone residual is called the N scheme (N stands for narrow), and it is due to P.L. Roe in the $\mathbb{P}^1$ case. The construction is as follows. We notice that the total residual on $\mathcal{K}$, thanks to the Gauss formula, also writes

$$\Phi^{\mathcal{K}}=\int_{\mathcal{K}}\text{div }\mathbf{f}(u^h)=\int_{\mathcal{K}}\nabla\mathbf{f}_u(u^h)\cdot\nabla u^h=\sum_{\sigma\in\mathcal{K}}\Big(\int_{\mathcal{K}}\nabla\mathbf{f}_u(u^h)\cdot\nabla\varphi_\sigma\Big)u_\sigma$$

We introduce the "inflow" parameters $k_\sigma=\int_{\mathcal{K}}\nabla\mathbf{f}_u(u^h)\cdot\nabla\varphi_\sigma)$, so that $\Phi^{\mathcal{K}}=\sum_\sigma k_\sigma u_\sigma$. We notice that $\sum_\sigma k_\sigma=0$. This parameters are called the inflow parameters because in the $\mathbb{P}^1$ case and for a linear flux, their sign characterizes whether the flow $\nabla_u\mathbf{f}(u^h)$ is inflow or outflow in the element $\mathcal{K}$. The N-scheme is then defined by

$$\Phi_\sigma^N=\max(k_\sigma,0)\big(u_\sigma-\overline{u}\big) \tag{20a}$$

$$\overline{u}=N\left(\sum_{\sigma\in\mathcal{K}}\min(k_\sigma,0)u_\sigma\right) \tag{20b}$$

$$N^{-1}=\sum_{\sigma\in\mathcal{K}}\min(k_\sigma,0) \tag{20c}$$

The average $\overline{u}$ is defined such that the relations (8) hold true. An easy calculation shows that

$$c_{\sigma'\sigma}^N=\min(k_\sigma,0)N\max(k_\sigma,0)\geq 0$$

so that the scheme is monotonicity preserving. Numerical experiments shows that this a very good first order for $\mathbb{P}^1$ element (hence for triangles and tetrahedrons) and provides less good results for higher elements.

Similarly, one can define an upwind high order scheme, nicknamed as the LDA scheme (Low Diffusion A schemes, there has been a LDB, less successful), it is defined by:

$$\Phi_\sigma^{LDA} = -\max(k_\sigma, 0) N\Phi.$$

It is a very good scheme for triangular/tet $\mathbb{P}^1$ elements, but it reveals to be unstable for higher elements or non triangular elements.

3.1. Explicit construction. The construction is local to an element (or boundary edge) $\mathcal{K}$, so we drop the dependency with respect to the element. We start from a monotone first order scheme, such as the Rusanov or the N scheme, denote the first order residuals in the element as $\{\Phi_\sigma^M\}_{\sigma\in K}$ and the high order residuals (to be constructed) by $\{\Phi_\sigma^H\}_\sigma$. We then make the following formal observation:

$$\text{for all } \sigma \in \mathcal{K}, \Phi_\sigma^H = \frac{\Phi_\sigma^H}{\Phi_\sigma^M}\Phi_\sigma^M,$$

so that if $\Phi_\sigma^M = \sum_{\sigma'\in\mathcal{K}} c_{\sigma\sigma'}^M(u_{\sigma'} - u_\sigma)$, we have

$$\begin{aligned}\phi_\sigma^H &= \frac{\Phi_\sigma^H}{\Phi_\sigma^M}\Big(\sum_{\sigma'\in\mathcal{K}} c_{\sigma\sigma'}^M(u_{\sigma'} - u_\sigma)\Big)\\ &= \sum_{\sigma'\in\mathcal{K}}\left(\frac{\Phi_\sigma^H}{\Phi_\sigma^M}c_{\sigma'\sigma}^M\right)(u_{\sigma'} - u_\sigma)\Big)\\ &= \sum_{\sigma'\in\mathcal{K}} c_{\sigma'\sigma}^H(u_{\sigma'} - u_\sigma)\Big)\end{aligned}$$

with $c_{\sigma'\sigma}^H := \dfrac{\Phi_\sigma^H}{\Phi_\sigma^M}c_{\sigma'\sigma}^M$. Hence, to have $c_{\sigma'\sigma}^H \geq 0$, it is enough that

$$\Phi_\sigma^H\ \Phi_\sigma^M \geq 0$$

Introducing the parameters $\beta_\sigma^M = \frac{\Phi_\sigma^M}{\Phi}$ and $\beta_\sigma^H = \frac{\Phi_\sigma^H}{\Phi}$ where Φ is the total residual on the element $\mathcal{K}$, we see that:

- $\Phi_\sigma^H\ \Phi_\sigma^M \geq 0$ is equivalent to $\beta_\sigma^M\ \beta_\sigma^H \geq 0$,
- the conservation relations translates into:

$$\sum_{\sigma\in\mathcal{K}} \beta_\sigma^M = \sum_{\sigma\in\mathcal{K}} \beta_\sigma^H = 1. \tag{21}$$

- In order to guaranty the condition (17), a sufficient condition is that : for any C, and u^h such that $||u^h||_\infty \leq C$, there exists C' such that $|\beta_\sigma^H| \leq C'(C)$, uniformly for all meshes $\mathcal{T}_h$.

These constraints can easily be interpreted geometrically. Consider an simplex $\mathcal{S} = (\mathbf{a}_1, \dots \mathbf{a}_{n_\mathcal{K}})$ of dimension $n_\mathcal{K} - 1$ points, i.e. a triangle when $n_\mathcal{K} = 3$, a tetrahedron for $n_\mathcal{K} = 4$ and so on. These points have nothing to do with the mesh, they are only used to represent easily the constraint (21): it is well known that any point $\mathbf{M}$ of an affine space

of dimension $n_{\mathcal{K}}-1$ can be uniquely described n in term of its barycentric coordinates with respect to $\mathcal{S}$:

$$M=\sum_{i=1}^{n_{\mathcal{K}}-1}\lambda_i\mathbf{a}_i,\qquad \sum_{i=1}^{n_{\mathcal{K}}-1}\lambda_i=1$$

so this suggest to interpret the parameters β_σ^M and β_σ^H as barycentric coordinates with respect to the simplex $\mathcal{S}$: we interpret a scheme as a point in this abstract affine space, and finding the mapping $(\beta_\sigma^M)_{\sigma\in\mathcal{K}}\mapsto(\beta_\sigma^H)_{\sigma\in\mathcal{K}}$ can be interpreted as to find a mapping from this affine space onto itself. Then, to make the discussion more visual, we switch to $n_{\mathcal{K}}=3$, see figure 1. The conditions $\beta_\sigma^H\beta_\sigma^L\geq 0$ are interpreted as saying that β_i^H and β_i^L must be on the same side of the line $\lambda_i=0$.

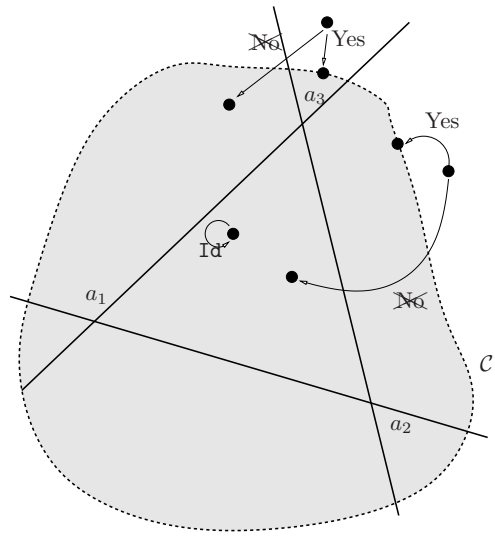


Figure 1. Geometrical representation of the monotonicity conditions. The invariant domain is materialized by the domain inside of $\mathcal{C}$.

The condition $|\beta_\sigma|\leq C$ is materialized, on figure (1), by the domain inside curve $\mathcal{C}$. Inside the invariant domain bounded by $\mathcal{C}$, the mapping is the identity, outside of $\mathcal{C}$ project the point $L=\sum_\sigma\beta_\sigma^L\mathbf{a}_\sigma$ on $\mathcal{C}$ without crossing the lines $\lambda_{\sigma_i}=0$. Once the β_σ^H are defined, we set simply $\Phi_\sigma^H=\beta_\sigma^H\Phi$.

The simplest invariant domain is certainly the simplex $(\mathbf{a}_1,\ldots,\mathbf{a}_{n_{\mathcal{K}}})$ for which $0\leq\lambda_\sigma\leq 1$. In that case, the most common formula is [6, 19]:

$$\beta_\sigma^H=\frac{\max(\beta_\sigma^M,0)}{\sum_{\sigma\in\mathcal{K}}\max(\beta_\sigma^M,0)}. \tag{22}$$

Note that $\sum_{\sigma\in\mathcal{K}}\max(\beta_\sigma^M,0)\geq 1$ because

$$1=\sum_{\sigma\in\mathcal{K}}\beta_\sigma^M=\sum_{\sigma\in\mathcal{K}}\max(\beta_\sigma^M,0)+\sum_{\sigma\in\mathcal{K}}\min(\beta_\sigma^M,0)\leq\sum_{\sigma\in\mathcal{K}}\max(\beta_\sigma^M,0).$$

When $\Phi = 0$, we simply set $\Phi_\sigma^H = 0$

In practice, this method is excellent for computing discontinuous solutions. When computing smoother solutions, we can see "wiggles" appearing, see section 5. They are not a manifestation of any instability since the scheme is perfectly L^∞ stable, but it is too over compressive, i.e. not dissipative enough.

It is quite easy to understand what is going on. We first, let us consider the problem on $[0,1]^2$:

$$\frac{\partial u}{\partial x} = 0 \tag{23}$$

with the boundary condition $u = g$ on $\{0\} \times [0,1]$. The grid is made of quadrangles, with vertexes (x_i, y_j), $x_i = \frac{i}{N}$, $y_j = \frac{j}{N}$, $0 \le i,j \le N$. The function g is piecewise linear, and $g(0,y_j) = (-1)^j$. The exact solution is independent of x.

The scheme is defined by

$$u_{ij}^{n+1} = u_{ij}^n - \omega_{ij} \sum_{K \ni (x_i,y_j)} \Phi_{i,j}^{H,K}(u_h^n)$$

with u_{ij}^0 given, and $u_{0j}^n = g(0,y_j)$. There are many ways of initializing, we consider two initializations:

- Initialization with the exact solution: $u_{ij}^0 = g(0,y_j) = (-1)^j$
- Check-board mode: $u_{ij}^0 = (-1)^{i+j}$

The solution at the n-th iteration is reconstructed with the $\mathbb{Q}^1$ interpolation. It is easy to see that for both initialization, we have, for any K,

$$\Phi^K = \int_{\partial K} u^h \mathbf{n}_x = 0$$

so that in both cases, for any i,j,n, $u_{ij}^n = u_{ij}^0$! The method, as it is, is not well posed, and there are spurious modes.

To remedy to this serious drawback, there are several possibilities, see [2]. The most flexible one is to add a streamline diffusion term:

$$\Phi_\sigma^{H,K,\star} = \Phi_\sigma^{H,K} + \theta_K h_K \int_K \left(\nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_\sigma\right) N \left(\nabla_u \mathbf{f}(u^h) \cdot \nabla u^h\right) \tag{24}$$

where N is define by (20b), and $\theta_K \approx 0$ in discontinuities and $\theta_K \approx 1$ away from discontinuities. When we apply this correction (with $\theta = 1$) to (23) this corrects the problem.

To see what is the rational behind (24), let us first switch to the one dimensional problem:

$$\begin{aligned} &\frac{\partial f(u)}{\partial x} = 0 \quad x \in [0,1] \\ &u(0) = u_0 \\ &u(1) = u_1. \end{aligned} \tag{25}$$

The boundary conditions are imposed weakly, and to make things simple, assume $f'(u_0) > 0$ and $f'(u_1) < 0$ so that the solution is $u = u_0$. The interval $[0,1]$ is discretized with the mesh

which elements are $[x_i, x_{i+1}]$, $0 = x_0 < x_1 < \ldots < x_{n-1}, x_n = 1$. Whatever the order, the total residual is for $K_{i+1/2} = [x_i, x_{i+1}]$

$$\Phi^{K_i+1/2} = f(u_{i+1}) - f(u_i)$$

so that the high order residuals are simply, for any degree of freedom $\sigma \in K$, $\Phi_\sigma^K = \beta_\sigma^K\big(f(u_{i+1} - f(u_i)\big)$. In particular, the internal degrees of freedom play no role. Assume now that $k = 1$, there is no internal degree of freedom, and let us evaluate the entropy balance for the entropy $U(u) = \frac{1}{2}u^2$:

$$\begin{aligned}\mathcal{E} &= \sum_{i=0}^{N-1} u_i\left(\beta_i^{K_{i-1/2}}\big(f(u_{i+1}) - f(u_i)\big) + \beta_i^{K_{i+1/2}}\big(f(u_{i+1}) - f(u_i)\big)\right)\\ &= \int_0^1 u^h \frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1}\left(\gamma_i^{K_{i+1/2}} u_i + \gamma_{i+1}^{K_{i+1/2}} u_{i+1/2}\right)\big(f(u_{i+1}) - f(u_i)\big)\end{aligned}$$

with $\gamma_j^{K_{i+1/2}} = \beta_j^{K_{i+1/2}} - \frac{1}{2}$

$$= \int_0^1 u^h \frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1} \gamma_{i+1}^{K_{i+1/2}}(f(u_{i+1}) - f(u_i))(u_{i+1} - u_i).$$

For the scheme to be dissipative, a sufficient condition is that for all i,

$$\gamma_{i+1}^{K_{i+1/2}}(f(u_{i+1}) - f(u_i))(u_{i+1} - u_i) \geq 0,$$

i.e.

$$\gamma_{i+1}^{K_{i+1/2}} \frac{f(u_{i+1}) - f(u_i)}{u_{i+1} - u_i} \geq 0$$

with a strict inequality for at least one interval.

The evaluation of $\beta_\sigma^{K_{i+1/2}}$ is done with the only aim of having an L^∞ stable scheme, so that this inequality might not be true [1]. Adding the streamline term, i.e.

$$\theta(u_{i+1} - u_i)\int_{x_i}^{x_{i+1}} N\big(\frac{\partial f}{\partial u}\big)^2 \frac{\partial \varphi_\sigma}{\partial x} = (u_{i+1} - u_i)\big|\frac{\partial f}{\partial u}\big|\big(\varphi_\sigma(x_{i+1}) - \varphi_\sigma(x_i)\big)$$

will modify the entropy into

$$\mathcal{E} = \int_0^1 u^h \frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1}\left(\gamma_{i+1}^{K_{i+1/2}} \frac{f(u_{i+1}) - f(u_i)}{u_{i+1} - u_i} + \theta\big|\frac{\partial f}{\partial u}\big|\right)(u_{i+1} - u_i)^2$$

and $\mathcal{E} \leq \int_0^1 u^h \frac{\partial f}{\partial x}(u^h)$ provided that $\theta \geq 1$.

In the general case, we have the following result:

Proposition 3.2. *There exists $\theta > 0$ which depends only on the polynomial degree r such that if $\hat{f}_{\mathbf{n}}$ is an E-flux, then* (12) *is true with the residuals defined by* (24)

[1] However, in 1D it is very simple to show that the sign condition is true, let us ignore this fact however.

Proof. We need to check (12). On the elements K, we get:

$$\begin{aligned}
&\sum_{\sigma\in K} u_\sigma \Big(\Phi_\sigma^{H,K} + \theta_K h_K \int_K \big(\nabla \mathbf{f}(u^h)\nabla u^h\big) N\big(\nabla \mathbf{f}(u^h)\nabla u^h\big)\Big) \\
&\quad = \int_{\partial K} \mathbf{g_n} + \sum_\sigma \gamma_\sigma^K (u_\sigma - u_{\sigma_1}) \int_K \operatorname{div} \mathbf{f}(u^h) \\
&\qquad + \theta_K h_K \int_K \big(\nabla \mathbf{f}(u^h)\nabla u^h\big) N\big(\nabla \mathbf{f}(u^h)\nabla u^h\big)\Big) \qquad (26)\\
&\quad = \int_{\partial K} \mathbf{g_n} + \Big(\sum_{\sigma\in K} \gamma_\sigma^K (u_\sigma - u_{\sigma_1})\Big) \int_K \nabla_u \mathbf{f}(u^h)\cdot \nabla u^h \\
&\qquad + \theta h_K \int_K \big(\nabla \mathbf{f}(u^h)\cdot u^h\big)^2 N.
\end{aligned}$$

We see that the second term of the last line can be written as :

$$\Big(u_{\sigma_2} - u_{\sigma_1}, \dots, u_{\sigma_{n_K}} - u_{\sigma_1}\Big)\big(M + \theta_K Q\big) \begin{pmatrix} u_{\sigma_2} - u_{\sigma_1} \\ \vdots \\ u_{\sigma_{n_K}} - u_{\sigma_1} \end{pmatrix}$$

with

$$\mathcal{E}_K = M_{\sigma\sigma'} = \gamma_\sigma^K \int_K \nabla_u f(u^h)\cdot \nabla \varphi_{\sigma'}$$

and

$$Q_{\sigma\sigma'} = h_K \int_K (\nabla_u \mathbf{f}(u^h)\nabla\varphi_\sigma) N (\nabla_u \mathbf{f}(u^h)\nabla\varphi_{\sigma'}).$$

The matrix Q is positive, $\ker Q \subset \ker M$. Since N is constant, we see that

$$\begin{aligned}
&\Big(u_{\sigma_2} - u_{\sigma_1}, \dots, u_{\sigma_{n_K}} - u_{\sigma_1}\Big) M \begin{pmatrix} u_{\sigma_1} - u_{\sigma_1} \\ \vdots \\ u_{\sigma_{n_K}} - u_{\sigma_1} \end{pmatrix} \\
&\geq -\sqrt{|K|} h_K \sqrt{\sum_\sigma (\gamma_\sigma^K)^2}\, \max_K ||\nabla u^h|| \sqrt{\int_K (\nabla_u \mathbf{f}(u^h)\cdot\nabla u^h)^2}.
\end{aligned}$$

Since $\mathbb{P}^r(K)$ is finite dimensional, there exists $C_{2,\infty}$ which depends only on r such that

$$\sqrt{|K|}\, \max_K ||\nabla u^h|| \leq C_{2,\infty} \sqrt{\int_K (\nabla u^h)^2}$$

so that

$$\begin{aligned}
\mathcal{E}_K \geq & - h_K C_{2,\infty} \sqrt{\sum_\sigma (\gamma_\sigma^K)^2} \sqrt{\int_K (\nabla u^h)^2} \sqrt{\int_K (\nabla_u \mathbf{f}(u^h)\cdot\nabla u^h)^2} \\
& + h_K \theta N \int_K (\nabla_u \mathbf{f}(u^h)\cdot \nabla u^h)^2.
\end{aligned}$$

The last thing to show is the existence of $C_r > 0$ such that $\sqrt{\int_K (\nabla_u \mathbf{f}(u^h)\cdot\nabla u^h)^2} \geq C\sqrt{\int_K (\nabla u^h)^2}$ on $\mathbb{P}_r(K)$. Since $\mathbb{P}_r(K) = \ker Q \bigoplus H$ where the two spaces are orthogonal with respect to the scalar product[2] $a(u,v) = \int_K \nabla u\cdot\nabla v$, and because the space are finite dimensional, there exists $C>0$ such that

$$\forall u \in U_h,\ \frac{\int_K (\nabla_u \mathbf{f}(u^h)\cdot\nabla u^h)^2}{\int_K (\nabla u^h)^2} \geq C_r > 0.$$

Connecting all the pieces together, since $\beta_\sigma^K \in [0,1]$. $\sum\sum_\sigma (\gamma_\sigma^K)^2 \leq n_K$, we see that $\theta \geq \frac{C_r}{C_{2,\infty}}$ guaranties the entropy inequality.

On the boundary element, if one takes and E-flux, the inequality is also valid. □

Remark 3.3. In practical simulations, $\theta = \frac{1}{n_K}$ is fine.

4. A variational formulation for RD schemes

Though described only by discrete formula, it is possible to identify the mapping χ in (5). Using the same technique, we see that

$$\begin{aligned}\Phi_\sigma^K =& \beta_\sigma^K \int_K \text{div } \mathbf{f}(u^h) + \theta h_K \int_K \big(\nabla_u \mathbf{f}(u^h)\nabla\varphi_\sigma\big) N\big(\nabla_u \mathbf{f}(u^h)\nabla u^h\big)\\ =& \int_K \chi_{u^h}(\varphi_\sigma)\ \text{div } \mathbf{f}(u^h),\end{aligned}$$

with

$$\chi_{u^h}(v^h) = \sum_{\sigma\in K}\left(\beta_\sigma^K v_\sigma + \theta h_K\big(\nabla_u \mathbf{f}(u^h)\nabla\varphi_\sigma\big)N\right). \tag{27}$$

5. Numerical examples

In this section, we illustrate the behavior of the method on two examples: a linear transport problem and a non linear one. In $\Omega = [0,1]^2$, we consider

$$\vec{\lambda} = (y,-x)^T \text{ and } u(x,y) = \varphi_0(x) \text{ if } y = 0 \tag{28}$$

with the boundary conditions

$$\varphi_0(x) = \begin{cases} \cos^2(2\pi x) & \text{if } x\in[\frac14,\frac34] \\ 0 & \text{else}\end{cases}$$

The isolines of the exact solution are circles of center $(0,0)$. The form of the Burgers equation is the following:

$$\begin{aligned}&\frac{\partial u}{\partial y} + \frac12\frac{\partial u^2}{\partial x} = 0 \qquad \text{if } x\in[0,1]^2\\ &u(x,y) = 1.5 - 2x \quad \text{on the inflow boundary.}\end{aligned} \tag{29a}$$

[2] if we remove the subspace of constant polynomial, which is included in $\ker Q$, this becomes a scalar product, thus sum is direct and H depends intrinsically on $\ker Q$.

The exact solution consists in a fan that merges into a shock which foot is located at $(x, y) = (3/4, 1/2)$. More precisely, the exact solution is

$$u(x,y) = \begin{cases} \text{if } y \geq 0.5 & \begin{cases} -0.5 & \text{if } -2(x-3/4)+(y-1/2) \geq 0 \\ 1.5 & \text{else} \end{cases} \\ \text{else} & \max\left(-0.5, \min\left(1.5, \dfrac{x-3/4}{y-1/2}\right)\right) \end{cases} \tag{29b}$$

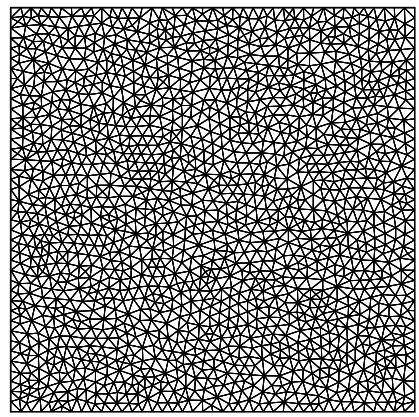


Figure 2. Mesh for the numerical experiments.

The mesh displayed on figure 2 is used to obtain the solutions shown on figure 3 and 4.

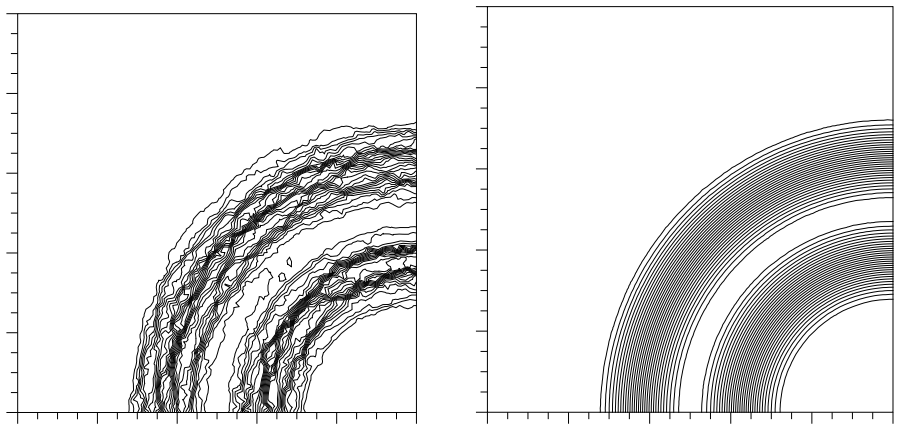
We see, on figure 3-(a) that without the streamline term in (24), the solution looks very wiggly. Again, it is not an instability, only a manifestation of spurious modes that are completely eliminated using (24). If one makes a convergence study on this problem using $\mathbb{P}^1$, $\mathbb{P}^2$ and $\mathbb{P}^3$ elements, we recover the expected order of convergence.

h	$\epsilon_{L^2}(\mathbb{P}^1)$	$\epsilon_{L^2}(\mathbb{P}^2)$	$\epsilon_{L^2}(\mathbb{P}^3)$
1/25	0.50493E-02	0.32612E-04	0.12071E-05
1/50	0.14684E-02	0.48741E-05	0.90642E-07
1/75	0.74684E-03	0.13334E-05	0.16245E-07
1/100	0.41019E-03	0.66019E-06	0.53860E-08
	$\mathcal{O}^{\text{ls}}_{L^2}$ =1.790	$\mathcal{O}^{\text{ls}}_{L^2}$ =2.848	$\mathcal{O}^{\text{ls}}_{L^2}$ =3.920

Table 1. Order of accuracy on refined mesh constructed from the mesh of figure 2, L^2 norm. The slopes are obtained by least square

Strictly speaking, the streamline in (24) destroys the maximum preserving nature of the scheme: the operator defined by (24) is not, a priori, of the type (18) with positive coefficients. We have not been able, so far, to analyze in full detail the scheme from this point of view, but all the numerical experiments that we have done so far, including with system case, indicate that the streamline term (24) acts as a filter, and does not spoil the monotonic-

ity preserving properties. Actually, this property is violated, but the over- and undershoot are negligible, as what occurs for the ENO and WENO schemes.



(a):without streamline term in (24) (b): with the streamline term in (24)

Figure 3. Solution of (28) with (22) and (24), $\mathbb{P}^2$ elements

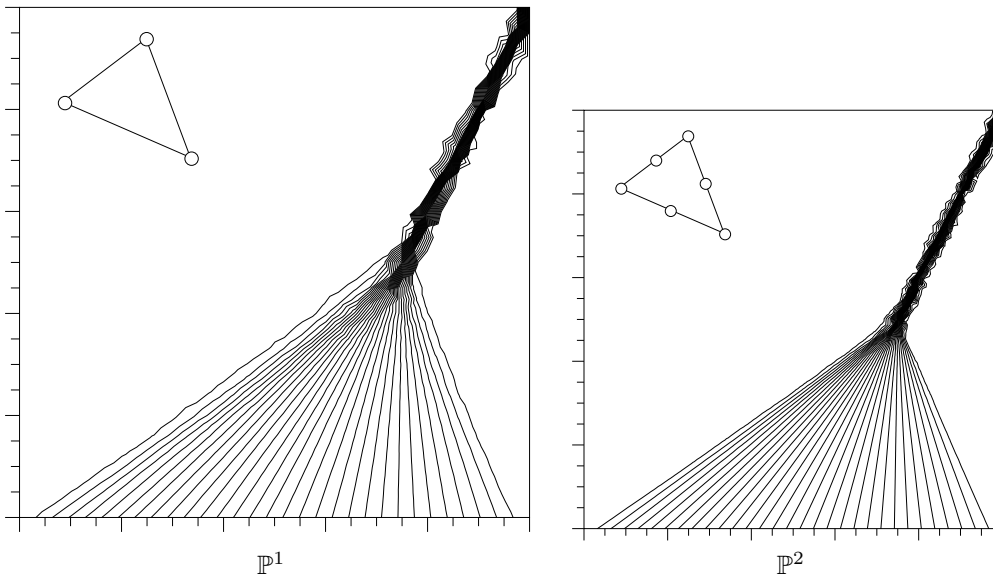


Figure 4. Solution of (29) with (24)

6. Flux formulation of Residual Distribution schemes

In this section we show that the scheme (7) also admits a flux formulation, with an explicit form of the flux. Hence the method is also locally conservative. This is well known for the Finite Volume and Discontinuous Galerkin approximation, much less understood for the RDS and continuous finite elements, despite the paper [12].

Let us consider any common edge or face Γ of K^+ and K^-, two elements. Let $\mathbf{n}$ be the normal to Γ, see Figure 5. A flux $\hat{\mathbf{f}}(S^+, S^-, \mathbf{n})$ between K^+ and K^- has to satisfy

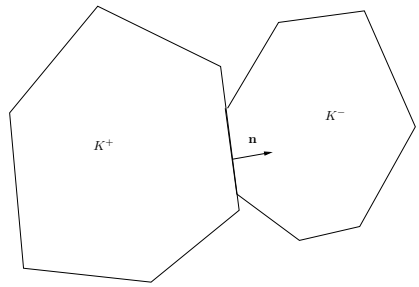


Figure 5.

$$F(S^+, S^-, \mathbf{n}) = -F(S^-, S^+, \mathbf{n}). \tag{30a}$$

and the consistency condition

$$F(S, S, \mathbf{n}) = f(S) \cdot \mathbf{n}. \tag{30b}$$

In (30a), the symbols $S^\pm$ represent set of states, where S^+ is associated to K^+ and S^- to K^-. For a first order finite volume scheme, we have $S^+ = \mathbf{u}_{K_+}$ and $S^- = \mathbf{u}_{K_-}$, the average values of $\mathbf{u}$ in K^+ and K^-. For the other schemes the definition is more involved. The aim of this section is to define $\hat{\mathbf{f}}$ and $S^\pm$ in the RDS case.

We briefly recall finite volume schemes. Then we show that RDS can be interpreted as finite volume schemes. To make the exposure easier, we assume that $d = 2$ and that the tessellation is conformal, made of triangles. This is not essential as the analysis shows it.

6.1. Analysis.

6.1.1. A recap on Finite volume methods. We denote the list of edges/faces of the elements of $\mathcal{K}$ by $\mathcal{G}$. Considering a numerical flux $\hat{\mathbf{f}}$, and a cell K, the formulation is

$$\int_{\partial K} \mathbf{f}(\mathbf{u}) \cdot \mathbf{n} \approx \sum_{\Gamma \in \mathcal{G}} \hat{\mathbf{f}}(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}_\Gamma)$$

so that an approximation of (1) is

$$|K| \frac{u_K^{n+1} - u_K^n}{\Delta t} + \sum_{\Gamma \in \mathcal{G}} F(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}) + \sum_{\Gamma \in \mathcal{G}, \Gamma \subset \partial\Omega^+} F(\mathbf{u}^+, \mathbf{g}, \mathbf{n}) = 0 \tag{31a}$$

and initial conditions

$$u_K^0 = \frac{\int_K u_0(x)dx}{|K|}. \tag{31b}$$

In (31a), we have specialized for the MUSCL method however this is not essential. We have chosen a simple Euler forward time stepping, more accurate solutions can be obtained using the method of lines, for example by using SSP Runge Kutta approximations [10]. More details can be found in [9, 15].

6.1.2. Finite volume as Residual distribution schemes. Here, we rephrase [1]. The notations are defined in Figure 6. Again, we specialize ourself to the case of triangular elements,

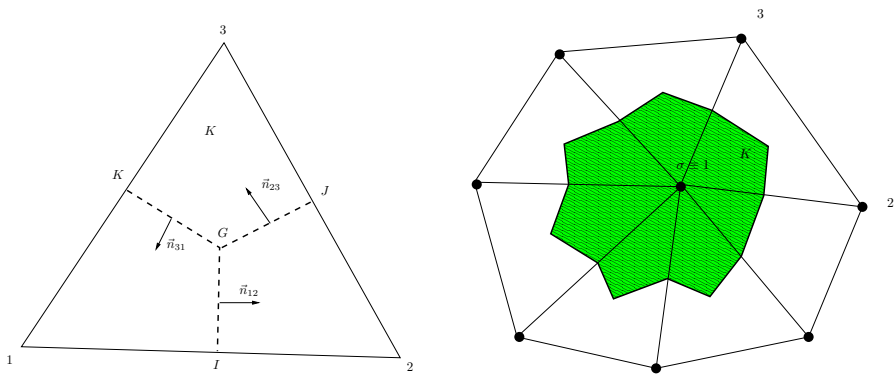


Figure 6. Notations for the finite volume schemes. On the left: definition of the control volume for the degree of freedom σ. The vertex σ plays the role of the vertex 1 on the left picture, etc for the triangle K.

but clearly *exactly the same arguments* can be given for more general elements, provided a conformal approximation space of the type U_h can be constructed. This is clearly the case for triangle elements, and we can take $p = 1$.

The control volume in this case are defined as the median cell, see figure 6. We concentrate on the div $\mathbf{f}$ approximation. Since the boundary of C is a closed polygon, we have

$$\sum_{\gamma \subset \partial C} \mathbf{n}_\gamma = 0$$

where γ is any of the segment included in ∂C, such as IG on Figure 6. Hence

$$\begin{aligned}\sum_{\gamma\subset\partial C} \hat{\mathbf{f}}(u_\sigma^+ u^-, \mathbf{n}_\gamma) &= \sum_{\gamma\subset\partial C} \hat{\mathbf{f}}(u_\sigma^+ u^-, \mathbf{n}_\gamma) - \left(\sum_{\gamma\subset\partial C} \mathbf{n}_\gamma\right)\cdot \mathbf{f}(\mathbf{u}_h(\sigma)) \\ &= \sum_{K,\sigma\in K} \sum_{\text{internal boundaries around } \sigma} \left(\hat{\mathbf{f}}(u_\sigma^+ u^-, \mathbf{n}_\gamma) - \mathbf{f}(\mathbf{u}_h(\sigma))\cdot \mathbf{n}_\gamma\right)\end{aligned}$$

To make things explicit, in K, the internal boundaries are IG, JG and KG, and those around $\sigma \equiv 1$ are IG and KG. We set

$$\Phi_\sigma^K = \sum_{\text{internal boundaries around } \sigma} \left(\hat{\mathbf{f}}(u_\sigma^+ u^-, \mathbf{n}_\gamma) - \mathbf{f}(\mathbf{u}_h(\sigma))\cdot \mathbf{n}_\gamma\right). \tag{32}$$

If now we sum up these three quantities and get:

$$\begin{aligned}\sum_{\sigma\in K}\Phi_\sigma^K &= \Big(\hat{\mathbf{f}}(u_1^+,u_2^+,\mathbf{n}_{12})-\hat{\mathbf{f}}(u_1^+,u_3^+,\mathbf{n}_{13})-\mathbf{f}(\mathbf{u}_1)\cdot\mathbf{n}_{12}+\mathbf{f}(\mathbf{u}_1)\cdot\mathbf{n}_{31}\Big)\\ &+\Big(\hat{\mathbf{f}}(u_2^+,u_3^+,\mathbf{n}_{23})-\hat{\mathbf{f}}(u_2^+,u_1^+,\mathbf{n}_{12})+\mathbf{f}(\mathbf{u}_2)\cdot\mathbf{n}_{12}-\mathbf{f}(\mathbf{u}_2)\cdot\mathbf{n}_{23}\Big)\\ &+\Big(-\hat{\mathbf{f}}(u_3^+,u_2^+,\mathbf{n}_{23})+\hat{\mathbf{f}}(u_3^+,u_1^+,\mathbf{n}_{31})-\mathbf{f}(\mathbf{u}_3)\cdot\mathbf{n}_{23}+\mathbf{f}(\mathbf{u}_3)\cdot\mathbf{n}_{31}\Big)\\ &=\mathbf{f}(\mathbf{u}_1)\cdot\big(\mathbf{n}_{12}-\mathbf{n}_{31}\big)+\mathbf{f}(\mathbf{u}_2)\cdot\big(-\mathbf{n}_{23}+\mathbf{n}_{31}\big)+\mathbf{f}(\mathbf{u}_3)\cdot\big(\mathbf{n}_{31}-\mathbf{n}_{23}\big)\\ &=\mathbf{f}(\mathbf{u}_1)\cdot\frac{\mathbf{n}_1}{2}+\mathbf{f}(\mathbf{u}_2)\cdot\frac{\mathbf{n}_2}{2}+\mathbf{f}(\mathbf{u}_3)\cdot\frac{\mathbf{n}_3}{2}\end{aligned}$$

where $\mathbf{n}_j$ is the scaled inward normal of the edge opposite to vertex σ_j, i.e. twice the gradient of the $\mathbb{P}^1$ basis function φ_{σ_j} associated to this degree of freedom. Thus, we can reinterpret the sum as the boundary integral of the Lagrange interpolant of the flux. The finite volume scheme is then a residual distribution scheme with residual defined by (32) and a total residual defined by

$$\Phi^K := \int_{\partial K}\mathbf{f}^h\cdot\mathbf{n}, \qquad \mathbf{f}^h=\sum_{\sigma\in K}\mathbf{f}(\mathbf{u}_\sigma)\varphi_\sigma. \tag{33}$$

6.1.3. Residual distribution schemes as finite volume schemes.. Let K be a fixed triangle. We are given a set of residues $\{\Phi_\sigma^K\}_{\sigma\in K}$, our aim here is to define a flux function such that relations similar to (32) hold true. We show the method for $\mathbb{P}^1$ and $\mathbb{P}^2$ interpolant, more general cases can easily be handled in the same way.

Warm up: The $\mathbb{P}^1$ case. Let us begin with the $\mathbb{P}^1$ case: the degrees of freedom are the vertexes of K, and we consider a linear interpolation in K. The flux across ID in the direction $\mathbf{n}_{12}$ is denoted by $\hat{\mathbf{f}}_{\mathbf{n}_{12}}$ and the flux across IG in the direction $-\mathbf{n}_{12}$ is $\hat{\mathbf{f}}_{-\mathbf{n}_{12}}=-\hat{\mathbf{f}}_{\mathbf{n}_{12}}$ by definition. Using similar notations, we must satisfy

$$\begin{aligned}\Phi_1&=\hat{\mathbf{f}}_{\mathbf{n}_{12}}-\hat{\mathbf{f}}_{\mathbf{n}_{31}}-\mathbf{f}(\mathbf{u}_1)\cdot\frac{\mathbf{n}_1}{2}\\ \Phi_2&=-\hat{\mathbf{f}}_{\mathbf{n}_{12}}+\hat{\mathbf{f}}_{\mathbf{n}_{23}}-\mathbf{f}(\mathbf{u}_2)\cdot\frac{\mathbf{n}_2}{2}\\ \Phi_3&=-\hat{\mathbf{f}}_{\mathbf{n}_{23}}+\hat{\mathbf{f}}_{\mathbf{n}_{31}}-\mathbf{f}(\mathbf{u}_3)\cdot\frac{\mathbf{n}_3}{2}\end{aligned} \tag{34}$$

Clearly, there is a compatibility relation:

$$\Phi^K=\sum_\sigma\mathbf{f}(\mathbf{u}_\sigma)\cdot\nabla\varphi_\sigma. \tag{35}$$

We can rewrite (34) as a linear system:

$$\begin{pmatrix}\Phi_1+\mathbf{f}(\mathbf{u}_1)\cdot\frac{\mathbf{n}_1}{2}\\ \Phi_2+\mathbf{f}(\mathbf{u}_2)\cdot\frac{\mathbf{n}_2}{2}\\ \Phi_3+\mathbf{f}(\mathbf{u}_3)\cdot\frac{\mathbf{n}_3}{2}\end{pmatrix}=\begin{pmatrix}1&-1&0\\-1&0&1\\0&1&-1\end{pmatrix}\begin{pmatrix}\hat{\mathbf{f}}_{\mathbf{n}_{12}}\\ \hat{\mathbf{f}}_{\mathbf{n}_{31}}\\ \hat{\mathbf{f}}_{\mathbf{n}_{23}}\end{pmatrix}:=A\begin{pmatrix}\hat{\mathbf{f}}_{\mathbf{n}_{12}}\\ \hat{\mathbf{f}}_{\mathbf{n}_{31}}\\ \hat{\mathbf{f}}_{\mathbf{n}_{23}}\end{pmatrix}$$

The matrix A is not invertible but has rank 2. Since (35) is true, there exists one solution at least. We can find easily one sample solution.

Let us first set $\hat{\mathbf{f}}_{\mathbf{n}_{31}} = 0$. Then we get

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= \Phi_1 + \mathbf{f}(\mathbf{u}_1)\cdot\tfrac{\mathbf{n}_1}{2}\\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= \Phi_1 + \Phi_2 + \mathbf{f}(\mathbf{u}_2)\cdot\tfrac{\mathbf{n}_2}{2} + \mathbf{f}(\mathbf{u}_1)\cdot\tfrac{\mathbf{n}_1}{2}\\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= 0
\end{aligned}$$

Thanks to (35), this can be rewritten as

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= \Phi_1 + \mathbf{f}(\mathbf{u}_1)\cdot\tfrac{\mathbf{n}_1}{2}\\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= -\Phi_3 - \mathbf{f}(\mathbf{u}_3)\cdot\tfrac{\mathbf{n}_3}{2}\\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= 0
\end{aligned}$$

Then we set $\hat{\mathbf{f}}_{\mathbf{n}_{12}} = 0$, thus

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= 0\\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= \Phi_2 + \mathbf{f}(\mathbf{u}_2)\cdot\tfrac{\mathbf{n}_2}{2}\\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= -\Phi_1 + \mathbf{f}(\mathbf{u}_1)\cdot\tfrac{\mathbf{n}_1}{2}.
\end{aligned}$$

Last, we set $\hat{\mathbf{f}}_{\mathbf{n}_{23}} = 0$ and get

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= -\Phi_2 - \mathbf{f}(\mathbf{u}_2)\cdot\tfrac{\mathbf{n}_2}{2}\\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= 0\\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= \Phi_3 + \mathbf{f}(\mathbf{u}_3)\cdot\tfrac{\mathbf{n}_3}{2}
\end{aligned}$$

To have a symmetric formulation, it is enough to take the average,

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= \frac{\Phi_1 - \Phi_2}{3} + \frac{1}{6}\Big(\mathbf{f}(\mathbf{u}_1)\cdot\mathbf{n}_1 - \mathbf{f}(\mathbf{u}_2)\cdot\mathbf{n}_2\Big)\\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= -\frac{\Phi_2 - \Phi_3}{3} + \frac{1}{6}\Big(\mathbf{f}(\mathbf{u}_2)\cdot\mathbf{n}_2 - \mathbf{f}(\mathbf{u}_3)\cdot\mathbf{n}_3\Big)\\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= \frac{\Phi_3 - \Phi_1}{3} + \frac{1}{6}\Big(\mathbf{f}(\mathbf{u}_3)\cdot\mathbf{n}_3 - \mathbf{f}(\mathbf{u}_1)\cdot\mathbf{n}_2\Big)
\end{aligned}$$

or, by introducing $\Psi_i = \Phi_i - \mathbf{f}(\mathbf{u}_i)\cdot\frac{\mathbf{n}_i}{2}$,

$$\hat{\mathbf{f}}_{\mathbf{n}_{12}} = \frac{1}{3}\big(\Psi_1 - \Psi_2\big), \quad \hat{\mathbf{f}}_{\mathbf{n}_{23}} = \frac{1}{3}\big(\Psi_2 - \Psi_3\big), \quad \hat{\mathbf{f}}_{\mathbf{n}_{31}} = \frac{1}{3}\big(\Psi_3 - \Psi_1\big). \tag{36}$$

Let us check the consistency of the flux. We first have to adapt the notion of consistency. As recalled in the Introduction, two of the key arguments in the proof of the Lax-Wendroff theorem are related to the structure of the flux, for classical finite volume schemes. In [6], the proof is adapted to the case of Residual Distribution schemes. The property that stands for the consistency is that if in an element, all the states are identical, then the residuals are all vanishing. Hence, we will say that

Definition 6.1. A multidimensional flux

$$\hat{\mathbf{f}} := \hat{\mathbf{f}}(\mathbf{u}_1, \ldots, \mathbf{u}_n, \mathbf{n})$$

is consistent if, when $\mathbf{u}_1 = \mathbf{u}_2 = \ldots = \mathbf{u}_n = \mathbf{u}$ then

$$\hat{\mathbf{f}}(\mathbf{u}, \ldots, \mathbf{u}, \mathbf{n}) = \mathbf{f}(\mathbf{u}) \cdot \mathbf{n}.$$

Let us show that the flux (36) are consistent in that sense. If the three states are equal to $\mathbf{u}$, then we have

$$\hat{\mathbf{f}}_{\mathbf{n}_{12}} = \frac{1}{6}\mathbf{f}(\mathbf{u}) \cdot \left(\mathbf{n}_1 - \mathbf{n}_2\right), \quad \hat{\mathbf{f}}_{\mathbf{n}_{23}} = \frac{1}{6}\mathbf{f}(\mathbf{u}) \cdot \left(\mathbf{n}_2 + \mathbf{n}_3\right), \quad \hat{\mathbf{f}}_{\mathbf{n}_{31}} = \frac{1}{6}\mathbf{f}(\mathbf{u}) \cdot \left(\mathbf{n}_3 - \mathbf{n}_2\right)$$

By symmetry, we only consider the first relation. Using the notations of the figure 6, we see that $\mathbf{n}_1 - \mathbf{n}_2$ is the normal of $\vec{BC} - \vec{CA} = \vec{BC} + \vec{AC}$. Since G is the centroid of the triangle, we see that $\vec{GC} = (\vec{AC} + \vec{BC})/3$, and thus we get

$$\hat{\mathbf{f}}_{\mathbf{n}_{12}} = \mathbf{f}(\mathbf{u}) \cdot \mathbf{n}_{12}.$$

This ends the proof.

We can state a couple of general remarks:

1. In general, the residuals depends on more than 2 arguments. For stabilized finite element methods, or the non linear stable residual distribution schemes, see e.g. [5, 12, 19], the residuals depends on the three states of K. Thus the formula (36) shows that the flux on more than two states in contrast to the 1D case. In the Finite volume case however, the support of the flux function is generally larger than the three states of K, think for example of an ENO/WENO method, of a simpler MUSCL one.
2. The formula (36) are influenced by the form of the total residual (33). We show in the next paragraph how this can be generalized.
3. We have set at the beginning that $\hat{\mathbf{f}}_{\mathbf{n}_{ij}} = -\hat{\mathbf{f}}_{-\mathbf{n}_{ij}}$. The formula (36) are antisymmetric with respect to the indices, and then do respect the assumed equality.

The example of the $\mathbb{P}^2$ approximation and the more general case. We consider the set-up defined by Figure 7. The triangle is splitted first into 4 sub-triangles K_1, K_2, K_3 and K_4. From this sub-triangulation, we can construct a dual mesh as in the $\mathbb{P}^1$ case and we have represented the 6 sub-zones that are the intersection of the dual control volumes and the triangle K. Our notations are as follow: given any sub-triangle K_ξ, if γ_{ij} is intersection between two adjacent control volumes (associated to σ_i and σ_j vertices of K_ξ), the normal to γ_{ij} in the direction σ_i to σ_j is denoted by $\mathbf{n}_{ij}^\xi$. Similarly the flux across γ_{ij} is denoted $\hat{\mathbf{f}}_{ij}^\xi$.

Following the same method as in the $\mathbb{P}^1$ case, we set:

$$\begin{array}{lll}
\Phi_1 &= \hat{\mathbf{f}}_{14}^1 - \hat{\mathbf{f}}_{61}^1 & + \int_{\partial C_1 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_2 &= -\hat{\mathbf{f}}_{42}^2 + \hat{\mathbf{f}}_{25}^2 & + \int_{\partial C_2 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_3 &= -\hat{\mathbf{f}}_{53}^3 + \hat{\mathbf{f}}_{36}^3 & + \int_{\partial C_3 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_4 &= -\hat{\mathbf{f}}_{14}^1 + \left(\hat{\mathbf{f}}_{46}^1 - \hat{\mathbf{f}}_{64}^4\right) + \left(\hat{\mathbf{f}}_{45}^4 - \hat{\mathbf{f}}_{54}^2\right) + \hat{\mathbf{f}}_{42}^2 & + \int_{\partial C_4 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_5 &= -\hat{\mathbf{f}}_{25}^2 + \left(\hat{\mathbf{f}}_{54}^2 - \hat{\mathbf{f}}_{45}^4\right) + \left(\hat{\mathbf{f}}_{56}^4 - \hat{\mathbf{f}}_{65}^3\right) + \hat{\mathbf{f}}_{53}^3 & + \int_{\partial C_5 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_6 &= -\hat{\mathbf{f}}_{36}^3 + \left(\hat{\mathbf{f}}_{65}^3 - \hat{\mathbf{f}}_{56}^4\right) + \left(\hat{\mathbf{f}}_{64}^4 - \hat{\mathbf{f}}_{46}^1\right) + \hat{\mathbf{f}}_{61}^1 & + \int_{\partial C_6 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n}.
\end{array} \tag{37}$$

We can group the terms in (37) by sub-triangles, namely:

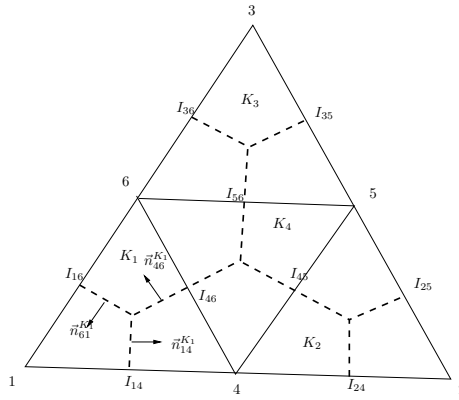


Figure 7. Geometrical elements for the $\mathbb{P}^2$ case.

$$
\begin{aligned}
\Phi_1 =& \left(\hat{\mathbf{f}}_{14}^1 - \hat{\mathbf{f}}_{61}^1 + \int_{\partial C_1 \cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
\Phi_2 =& \left(-\hat{\mathbf{f}}_{42}^2 + \hat{\mathbf{f}}_{25}^2 + \int_{\partial C_2 \cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
\Phi_3 =& \left(-\hat{\mathbf{f}}_{53}^3 + \hat{\mathbf{f}}_{36}^3 + \int_{\partial C_3 \cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
\Phi_4 =& \left(-\hat{\mathbf{f}}_{14}^1 + \hat{\mathbf{f}}_{46}^1 + \int_{\partial C_4 \cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}_{64}^4 + \hat{\mathbf{f}}_{45}^4 + \int_{\partial C_4 \cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
\Phi_5 =& \left(-\hat{\mathbf{f}}_{25}^2 + \hat{\mathbf{f}}_{54}^2 + \int_{\partial C_5 \cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}_{45}^4 + \hat{\mathbf{f}}_{56}^4 + \int_{\partial C_5 \cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}_{65}^3 + \hat{\mathbf{f}}_{53}^3 + \int_{\partial C_5 \cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
\Phi_6 =& \left(-\hat{\mathbf{f}}_{36}^3 + \hat{\mathbf{f}}_{65}^3 + \int_{\partial C_6 \cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}_{56}^4 + \hat{\mathbf{f}}_{64}^4 + \int_{\partial C_6 \cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}_{46}^1 + \hat{\mathbf{f}}_{61}^1 + \int_{\partial C_6 \cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n} \right)
\end{aligned}
\tag{38}
$$

where we have used:

$$\int_{\partial C_4\cap K}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}=\int_{\partial C_4\cap K_1}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}+\int_{\partial C_4\cap K_4}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}+\int_{\partial C_4\cap K_2}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$
$$\int_{\partial C_5\cap K}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}=\int_{\partial C_5\cap K_2}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}+\int_{\partial C_5\cap K_4}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}+\int_{\partial C_5\cap K_3}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$
$$\int_{\partial C_6\cap K_6}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}=\int_{\partial C_6\cap K_3}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}+\int_{\partial C_6\cap K_4}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}+\int_{\partial C_6\cap K_1}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$

Then we define the sub-residuals per sub elements:

$$\Phi_1^1=-\hat{\mathbf{f}}_{61}^1+\hat{\mathbf{f}}_{14}^1+\int_{\partial C_1\cap K_1}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n},\qquad \Phi_4^2=-\hat{\mathbf{f}}_{54}^2+\hat{\mathbf{f}}_{42}^2+\int_{\partial C_4\cap K_2}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$
$$\Phi_4^1=-\hat{\mathbf{f}}_{14}^1+\hat{\mathbf{f}}_{46}^1+\int_{\partial C_4\cap K_1}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n},\qquad \Phi_2^2=-\hat{\mathbf{f}}_{42}^2+\hat{\mathbf{f}}_{25}^2+\int_{\partial C_2\cap K_2}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$
$$\Phi_6^1=-\hat{\mathbf{f}}_{46}^1+\hat{\mathbf{f}}_{61}^1+\int_{\partial C_6\cap K_1}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n},\qquad \Phi_5^2=-\hat{\mathbf{f}}_{25}^2+\hat{\mathbf{f}}_{54}^2+\int_{\partial C_5\cap K_2}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$

$$\Phi_5^3=-\hat{\mathbf{f}}_{65}^3+\hat{\mathbf{f}}_{53}^3+\int_{\partial C_5\cap K_3}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n},\qquad \Phi_4^4=-\hat{\mathbf{f}}_{64}^4+\hat{\mathbf{f}}_{45}^4+\int_{\partial C_4\cap K_4}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$
$$\Phi_3^3=-\hat{\mathbf{f}}_{36}^3+\hat{\mathbf{f}}_{65}^3+\int_{\partial C_6\cap K_3}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n},\qquad \Phi_5^4=-\hat{\mathbf{f}}_{45}^4+\hat{\mathbf{f}}_{56}^4+\int_{\partial C_5\cap K_4}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$
$$\Phi_6^3=-\hat{\mathbf{f}}_{36}^3+\hat{\mathbf{f}}_{65}^3+\int_{\partial C_6\cap K_3}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n},\qquad \Phi_6^4=-\hat{\mathbf{f}}_{56}^4+\hat{\mathbf{f}}_{64}^4+\int_{\partial C_6\cap K_4}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}.\tag{39}$$

Clearly,

$$\Phi_1^1+\Phi_4^1+\Phi_6^1=\int_{\partial K_1}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n},\qquad \Phi_4^2+\Phi_2^2+\Phi_5^2=\int_{\partial K_2}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}$$
$$\Phi_5^3+\Phi_3^3+\Phi_6^3=\int_{\partial K_3}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n},\qquad \Phi_4^4+\Phi_5^4+\Phi_6^4=\int_{\partial K_4}\mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}\tag{40}$$

so we are back to the $\mathbb{P}^1$ case: in each sub-triangle, we can define flux that will depend on the 6 states of the element via the boundary flux. This is legitimate because in the $\mathbb{P}^1$ case, we have not used the fact that the interpolation is linear, we have only used the fact that we have 3 vertices. Clearly the flux are consistent in the sense of definition 6.1.

The same argument can be clearly extended to higher degree element, as well as to non triangular element: what is needed is to subdivide the element into sub-triangles.

Acknowledgments. The author has been partially supported by the FP7 ERC Advanced Grant ADDECCO # 226316 and the Swiss SNF grant # 200021_153604/1. I would like to thank Mario Ricchiuto, INRIA: our discussions have been very helpfull to simplify some of the arguments developped in this paper.

References

[1] R. Abgrall, *Toward the ultimate conservative scheme: Following the quest,* J. Comput. Phys., **167**(2) (2001), 277–315.

[2] _____, *Essentially non-oscillatory residual distribution schemes for hyperbolic problems,* J. Comput. Phys., **214**(2) (2006), 773–808.

[3] R. Abgrall and D. de Santis, *Linear and non-linear high order accurate residual distribution schemes for the discretization of the steady compressible navier-stokes equations*, J. Comput. Phys., 2014. in revision.

[4] R. Abgrall and D. de Santis, *High-order preserving residual distribution schemes for advection-diffusion scalar problems on arbitrary grids*, SIAM I. Sci. Comput., in press. also http://hal.inria.fr/docs/00/76/11/59/PDF/8157.pdf.

[5] R. Abgrall, A. Larat, and M. Ricchiuto, *Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid unstructured meshes,* J. Comput. Phys., 230(11) (2011), 4103–4136.

[6] R. Abgrall and P. L. Roe, *High-order fluctuation schemes on triangular meshes,* J. Sci. Comput., **19**(1-3) (2003), 3–36.

[7] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous finite element method for conservation laws IV: the multidimensional case*, Math. Comp., **54** (1990), 545–581.

[8] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta* local projection discontinuous Galerkin finite element method for conservation laws II: General framework. Math. Comp., **52** (1989), 411–435.

[9] E. Godlewski and P.-A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws.* New York, NY: Springer, 1996.

[10] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretisation methods*, SIAM Review, **43**(1) (2001), 89–112.

[11] A. Harten, *On a class of high resolution total-variational-stable finite-difference schemes (with appendix by Peter D. Lax).* SIAM J. Numer. Anal., **21** (1984), 1–23.

[12] T.J.R. Hughes, L.P. Franca, and M. Mallet, *A new finite element formulation for CFD: I*, symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. Comp. Meth. Appl. Mech. Engrg., **54** (1986), 223–234.

[13] C. Johnson, U. Nävert, and J. Pitkäranta, Finite element methods for linear hyperbolic problems. *Computer methods in applied mechanics and engineering*, **45** (1985), 285–312.

[14] D. Kröner, M. Rokyta, and M. Wierse, *A Lax-Wendroff type theorem for upwind finite volume schemes in* 2-*d*, East-West J. Numer. math., **4**(4) (1996), 279–292.

[15] R. J. Leveque, *Finite volume methods for hyperbolic problems,* Cambridge: Cambridge University Press, 2002.

[16] R.-H. Ni, A multiple grid scheme for solving the Euler equations. In *5th Computational Fluid Dynamics Conference*, pages 257–264, 1981.

[17] P.L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes,* J. Comput. Phys., **43** (1981), 357–372.

[18] _____, Characteristic-based schemes for the Euler equations. Annu. Rev. Fluid Mech. **18** (1986), 337–365.

[19] R. Struijs, H. Deconinck, and P.L. Roe, Fluctuation splitting schemes for the 2D Euler equations. VKI-LS 1991-01, 1991. Computational Fluid Dynamics.

Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich
E-mail: remi.abgrall@math.uzh.ch

Spline differential forms

Annalisa Buffa

Abstract. We introduce spline discretization of differential forms and study their properties. We analyse their geometric and topological structure, as related to the connectivity of the underlying mesh, we present degrees of freedom and we construct commuting projection operators, with optimal stability and approximation properties.

Mathematics Subject Classification (2010). 65N30, 65D07.

Keywords. Numerical analysis, spline theory, discretization of partial differential equations.

1. Introduction

This paper is a review of the work I have done mainly in collaboration with G. Sangalli and R. Vázquez on the definition and study of the spline approximation of differential forms, with the aim of writing spline-based numerical techniques for the solution of partial differential equations whose unknowns can be interpreted as differential forms. In this presentation I follow four main contributions of ours: [1–3] and the recent review paper [4].

The idea of using splines as basic tool for the discretization of partial differential equations traces back to the seventies but spline based methods never really became a standard practice due to several reasons: from the difficulty in setting boundary conditions, to the limitations imposed by their tensor product structure. Only in 2005, spline-based methods, together with the isoparametric paradigm, have been promoted in the mechanical engineering community under the name of isogeometric analysis, by T.J.R. Hughes and coauthors in the seminal paper [5] (see also the book [6]). Since then, spline-based (or isogeometric) methods have attracted a growing attention from the academic community.

The main challenge the authors of [5] wanted to address is to improve the interoperability between computer aided design systems (CAD) and partial differential equation (PDE) solvers and, for this reason, they have proposed to use CAD mathematical primitives, i.e. splines and NURBS ([7]), also to represent PDE unknowns. As unexpected consequence, it has been understood that the use of spline functions (or their generalizations), together with isoparametric concepts, results in an extremely successful idea and paves the way to many new numerical schemes enjoying features that would be extremely hard to achieve within a standard finite element framework.

In this paper, we focus our attention on the construction of the so-called spline complex, i.e., spline approximation spaces for the De Rham diagram. In the finite element context, the construction of discrete De Rham complexes has been object of intense study and its various aspects have been object of three review papers: for computational electromagnetics [8], for

Proceedings of the International Congress of Mathematicians, Seoul, 2014

finite element exterior calculus [9] and for eigenvalue problems [10]. It should be mentioned finite element exterior calculus was initiated in the seminal ICM paper [11].

Along this paper, we will work in three space dimensions however the constructions and results apply to any space dimensions with simple changes. The reason for this choice is to adopt the language of vector fields which may be closer to the finite element language than the language of differential forms.

In Section 2, we recall the main definition and we set our notation for spline spaces and projections. In Section 3 we briefly recall the De Rham diagram in a simple setting, in Section 4 we construct a subcomplex with spline spaces on the parametric domain $\widehat{\Omega} = (0,1)^3$. This is called *the spline complex*. We also construct commuting projectors and study the approximation properties that are obtained following the lines of [4].

Finally, we construct the spline complex on a general domain Ω that is supposed to be a union of several "patches" (i.e., spline mappings of the parametric domain), and analyse their approximation properties. This is the object of Section 5 and 6. Finally, we discuss our conclusions in Section 7.

2. Basics on splines

2.1. Definition and properties of univariate B-splines. Given two positive integers p and n, we say that $\Xi := \{\xi_1, \dots, \xi_{n+p+1}\}$ is a p-open knot vector if

$$\xi_1 = \dots = \xi_{p+1} < \xi_{p+2} \leq \dots \leq \xi_n < \xi_{n+1} = \dots = \xi_{n+p+1},$$

where repeated knots are allowed. Without loss of generality, we assume in the following that $\xi_1 = 0$ and $\xi_{n+p+1} = 1$.

We introduce also the vector $Z = \{\zeta_1, \dots, \zeta_N\}$ of knots without repetitions such that:

$$\Xi = \{\underbrace{\zeta_1, \dots, \zeta_1}_{m_1 \text{ times}}, \underbrace{\zeta_2, \dots, \zeta_2}_{m_2 \text{ times}}, \dots, \underbrace{\zeta_N, \dots, \zeta_N}_{m_N \text{ times}}\}, \tag{2.1}$$

with $\sum_{i=1}^{N} m_i = n + p + 1$, and $m_j \leq p + 1$ for all internal knots. Note that the points in Z form a partition of the unit interval $I = (0,1)$, i.e., a mesh, and the local mesh size of the element $I_i = (\zeta_i, \zeta_{i+1})$ is called $h_i = \zeta_{i+1} - \zeta_i$, for $i = 1, \dots, N-1$.

From the knot vector Ξ, B-spline functions of degree p are defined following the well-known Cox-DeBoor recursive formula: we start with piecewise constants ($p = 0$):

$$\widehat{B}_{i,0}(\zeta) = \begin{cases} 1 & \text{if } \xi_i \leq \zeta < \xi_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

and for $p \geq 1$ the *B-spline* functions are defined by the recursion

$$\widehat{B}_{i,p}(\zeta) = \frac{\zeta - \xi_i}{\xi_{i+p} - \xi_i} \widehat{B}_{i,p-1}(\zeta) + \frac{\xi_{i+p+1} - \zeta}{\xi_{i+p+1} - \xi_{i+1}} \widehat{B}_{i+1,p-1}(\zeta), \tag{2.3}$$

where it is here formally assumed that 0/0 = 0.

This gives a set of n B-splines that, among many other properties, are non-negative and form a partition of unity. They also form a basis of the space of *splines*, that is, piecewise polynomials of degree p with $k_j := p - m_j$ continuous derivatives at the points ζ_j, for

$j = 1, \ldots, N$. Therefore, $-1 \leq k_j \leq p-1$, and the maximum multiplicity allowed, $m_j = p+1$, gives $k_j = -1$, which stands for a discontinuity at ζ_j.

We denote the *univariate spline space* spanned by the B-splines by

$$S_p(\Xi) = \text{span}\{\widehat{B}_{i,p},\, i = 1, \ldots, n\}. \tag{2.4}$$

Note that the definition of each B-spline $\widehat{B}_{i,p}$ depends only on $p+2$ knots, which are collected in the *local knot vector* $\Xi_{i,p} := \{\xi_i, \ldots, \xi_{i+p+1}\}$ and, clearly $\text{supp}(\widehat{B}_{i,p}) = [\xi_i, \xi_{i+p+1}]$. Moreover, given an interval $I_j = (\zeta_j, \zeta_{j+1})$ of the partition, which can also be written as (ξ_i, ξ_{i+1}) for a certain (unique) i, we associate the *support extension* $\widetilde{I}_j$ defined as

$$\widetilde{I}_j := (\xi_{i-p}, \xi_{i+p+1}), \tag{2.5}$$

that is the interior of the union of the supports of basis functions whose support intersects I_j.

For the refinement of spline spaces, we call $h-$ refinement the insertion of new knots (see e.g., [12]), p-refinement the increase on the degree p while keeping the inter-regularity fixed, and k-refinement the one obtained by applying the Cox-DeBoor formula. Clearly both h and p refinement generates a sequence of embedded spaces and we say, in general, that $S_p(\Xi)$ is a *refinement* of $S_{p^0}(\Xi^0)$ if

$$S_{p^0}(\Xi^0) \subset S_p(\Xi). \tag{2.6}$$

Assuming the maximum multiplicity of the internal knots is less than or equal to the degree p, i.e., the B-spline functions are at least continuous, the derivative of each B-spline $\widehat{B}_{i,p}$ is given by the expression

$$\frac{d\widehat{B}_{i,p}}{d\zeta}(\zeta) = \frac{p}{\xi_{i+p} - \xi_i}\widehat{B}_{i,p-1}(\zeta) - \frac{p}{\xi_{i+p+1} - \xi_{i+1}}\widehat{B}_{i+1,p-1}(\zeta). \tag{2.7}$$

where we have assumed that $\widehat{B}_{1,p-1}(\zeta) = \widehat{B}_{n+1,p-1}(\zeta) = 0$. In fact, the derivative belongs to the spline space $S_{p-1}(\Xi')$, where $\Xi' = \{\xi_2, \ldots, \xi_{n+p}\}$ is a $(p-1)$-open knot vector. Moreover, it is easy to see that $\dfrac{d}{d\zeta} : S_p(\Xi) \to S_{p-1}(\Xi')$ is a surjective application. For later use, we define the so called *Curry-Schoenberg spline basis* (see e.g., [13, Ch. IX]), as follows

$$\widehat{D}_{i,p-1}(\zeta) = \frac{p}{\xi_{i+p+1} - \xi_{i+1}}\widehat{B}_{i+1,p-1}(\zeta), \quad \text{for } i = 1, \ldots, n-1.$$

The indices for the new basis have been shifted in order to start numbering from 1. Then formula (2.7) becomes

$$\frac{d\widehat{B}_{i,p}}{d\zeta}(\zeta) = \widehat{D}_{i-1,p-1}(\zeta) - \widehat{D}_{i,p-1}(\zeta), \tag{2.8}$$

where, again, we adopt the convention $\widehat{D}_{0,p-1} = \widehat{D}_{n,p-1} = 0$.

We end this section, recalling the concept of Greville sites. For each B-spline basis function $\widehat{B}_{j,p}$, we associate a *Greville site*, also called knot average:

$$\gamma_{j,p} = \frac{\xi_{j+1} + \ldots + \xi_{j+p}}{p}, \qquad j = 1, \ldots, n. \tag{2.9}$$

Greville sites are the coefficients of the identity in the expansion in B-spline basis, i.e.

$$\zeta = \sum_{j=1}^{n} \gamma_{j,p}\, \widehat{B}_{j,p}(\zeta). \tag{2.10}$$

Clearly, the Greville abscissa $\gamma_{j,p}$ depends only upon the local knot vector $\Xi_{j,p}$. When needed, we will adopt the notation $\gamma_{j,p} = \gamma[\Xi_{j,p}]$. It is immediate to see that, when the multiplicity of the internal knots is $m_j \leq p$, for $j = 2, \ldots, n-1$, the Greville points $\gamma_{j,p}$ are all distinct and form a partition of the interval $[0, 1]$, which we call *Greville mesh*.

2.2. Projections and quasi-interpolation operators. In this section we introduce interpolation and projection operators onto the space of splines $S_p(\Xi)$. There are several ways to define projections for splines, and here we only describe the ones that will be needed in the sequel of the paper.

In the present contribution we will often make use of the following local quasi-uniformity condition on the knot vector, that is a classical assumption in the mathematical isogeometric literature.

Assumption 2.1. *The partition defined by the knots $\zeta_1, \zeta_2, \ldots, \zeta_N$ is locally quasi-uniform, that is, there exists a constant $\theta \geq 1$ such that the mesh sizes $h_i = \zeta_{i+1} - \zeta_i$ satisfy the relation $\theta^{-1} \leq h_i/h_{i+1} \leq \theta$, for $i = 1, \ldots, N-2$.*

Since splines are not in general interpolatory, a common way to define projections is by defining a dual basis, i.e,

$$\Pi_{p,\Xi} : C^\infty([0,1]) \to S_p(\Xi), \qquad \Pi_{p,\Xi}(f) = \sum_{j=1}^{n} \lambda_{j,p}(f) \widehat{B}_{j,p}, \tag{2.11}$$

where $\lambda_{j,p}$ are a set of dual functionals verifying

$$\lambda_{j,p}(\widehat{B}_{k,p}) = \delta_{jk}, \tag{2.12}$$

δ_{jk} being the standard Kronecker symbol. It is trivial to prove that, thanks to this property, the quasi-interpolant $\Pi_{p,\Xi}$ preserves splines, that is

$$\Pi_{p,\Xi}(f) = f, \quad \forall f \in S_p(\Xi). \tag{2.13}$$

Following [12, Theorem 4.37], and [14], it holds the following.

Proposition 2.2. *There exists a choice of $\{\lambda_{j,p}\}_{j=1..n}$ such that, for any non empty knot span $I_i = (\zeta_i, \zeta_{i+1})$ it holds*

$$\|\Pi_{p,\Xi}(f)\|_{L^2(I_i)} \leq C \|f\|_{L^2(\widetilde{I}_i)}, \tag{2.14}$$

where the constant C depends only upon the degree p, and $\widetilde{I}_i$ is the support extension defined in (2.5)*. Moreover, if Assumption 2.1 holds, we also have*

$$|\Pi_{p,\Xi}(f)|_{H^1(I_i)} \leq C |f|_{H^1(\widetilde{I}_i)}, \tag{2.15}$$

with the constant C depending only on p and θ, and where H^1 denotes the Sobolev space of order one, endowed with the standard norm and seminorm.

The proof can be found in [4].

The operator $\Pi_{p,\Xi}$ can be modified in order to match boundary conditions. We can define, for all $f \in C^\infty([0,1])$:

$$\begin{aligned} &\widetilde{\Pi}_{p,\Xi}(f) = \sum_{j=1}^{n} \widetilde{\lambda}_{j,p}(f)\widehat{B}_{j,p} \quad \text{with} \\ &\widetilde{\lambda}_{1,p}(f) = f(0)\,,\ \widetilde{\lambda}_{n,p}(f) = f(1)\,,\ \widetilde{\lambda}_{j,p}(f) = \lambda_{j,p}(f), \quad j = 2,\dots,n-1. \end{aligned} \tag{2.16}$$

Clearly, the L^2 stability stated for $\Pi_{p,\Xi}$ cannot be valid for $\widetilde{\Pi}_{p,\Xi}$, but a similar result holds.

Proposition 2.3. *For any non empty knot span $I_i = (\zeta_i, \zeta_{i+1})$ it holds*

$$\|\widetilde{\Pi}_{p,\Xi}(f)\|_{L^2(I_i)} \le C\big(\|f\|_{L^2(\widetilde{I}_i)} + \widetilde{h}_i|f|_{H^1(\widetilde{I}_i)}\big) \tag{2.17}$$

where the constant C depends upon the degree p, and $\widetilde{I}_i$ is the support extension defined in (2.5)*, and $\widetilde{h}_i$ its length. Moreover, if Assumption 2.1 holds, we also have*

$$|\widetilde{\Pi}_{p,\Xi}(f)|_{H^1(I_i)} \le C\|f\|_{H^1(\widetilde{I}_i)} \tag{2.18}$$

with the constant C depending only on p, θ and where the space H^1 was already introduced in the previous proposition.

For the proof of this statement, we defer the reader to [4].

We end this section with the construction of another quasi-interpolant that will be useful later on, and concerns the construction of commuting projectors. In particular, given a projector $\Pi_{p,\Xi}$ constructed as above, we define

$$\Pi^c_{p-1,\Xi'} g\,(\zeta) := \frac{d}{d\zeta}\Pi_{p,\Xi} \int_0^\zeta g(s)\,\mathrm{d}s, \tag{2.19}$$

for all functions g such that $f(\zeta) = \int_0^\zeta g(s)\,\mathrm{d}s$ is in the domain of definition of $\Pi_{p,\Xi}$. The index c stands for commuting and it is indeed trivial to see that

$$\Pi^c_{p-1,\Xi'} \frac{d}{d\zeta} f = \frac{d}{d\zeta}\Pi_{p,\Xi} f \tag{2.20}$$

for all f in the domain of definition of $\Pi_{p,\Xi}$. Moreover, and as a consequence of the spline preserving property (2.13), it is also immediate to prove that $\Pi^c_{p-1,\Xi'}$ preserves B-splines, that is

$$\Pi^c_{p-1,\Xi'} g = g \qquad \forall g \in S_{p-1}(\Xi'). \tag{2.21}$$

Thus, we have the following commuting diagram

$$\begin{array}{ccccccc} \mathbb{R} & \longrightarrow & H^1(0,1) & \xrightarrow{\frac{d}{d\zeta}} & L^2(0,1) & \longrightarrow & 0 \\ & & \Big\downarrow{\scriptstyle \Pi_{p,\Xi}} & & \Big\downarrow{\scriptstyle \Pi^c_{p-1,\Xi'}} & & \\ \mathbb{R} & \longrightarrow & S_p(\Xi) & \xrightarrow{\frac{d}{d\zeta}} & S_{p-1}(\Xi') & \longrightarrow & 0. \end{array} \tag{2.22}$$

We prove the following proposition.

Proposition 2.4. *Let $g \in L^2(0,1)$, and let the projector $\Pi_{p,\Xi}$ be defined as in* (2.11)*, that is, $\Pi_{p,\Xi} f(\zeta) = \sum_{i=1}^{n} \lambda_{i,p}(f)\widehat{B}_{i,p}(\zeta)$ for any $f \in L^2(0,1)$. Then it holds:*

$$\Pi^c_{p-1,\Xi'} g(\zeta) = \sum_{j=1}^{n-1} \lambda^c_{j,p-1}(g)\widehat{D}_{j,p-1}(\zeta),$$

with

$$\lambda^c_{j,p-1}(g) = \lambda_{j+1,p}\left(\int_{\xi_j}^{\zeta} g(s)\mathrm{d}s\right) - \lambda_{j,p}\left(\int_{\xi_j}^{\zeta} g(s)\mathrm{d}s\right). \tag{2.23}$$

Moreover, if Assumption 2.1 is satisfied, then for all $I_i = (\zeta_i, \zeta_{i+1})$, it holds:

$$\|\Pi^c_{p-1,\Xi'} g\|_{L^2(I_i)} \le C\|g\|_{L^2(\widetilde{I}_i)}, \tag{2.24}$$

where $\widetilde{I}_i$ is the support extension of I_i defined in (2.5)*.*

Proof. Let $f(\zeta) := \int_0^{\zeta} g(s)\mathrm{d}s$. By definition of $\Pi^c_{p-1,\Xi'}$ and $\Pi_{p,\Xi}$, and then using the expression for the derivative (2.8), we have

$$\Pi^c_{p-1,\Xi'} g(\zeta) = \frac{d}{d\zeta}\Pi_{p,\Xi} f(\zeta) = \frac{d}{d\zeta}\sum_{i=1}^{n} \lambda_{i,p}(f)\widehat{B}_{i,p}(\zeta) = \sum_{i=1}^{n} \lambda_{i,p}(f)(\widehat{D}_{i-1,p}(\zeta) - \widehat{D}_{i,p}(\zeta)),$$

and recalling the convention $\widehat{D}_{0,p}(\zeta) = \widehat{D}_{n,p}(\zeta) = 0$, we obtain

$$\Pi^c_{p-1,\Xi'} g(\zeta) = \sum_{j=1}^{n-1} \left(\lambda_{j+1,p}(f) - \lambda_{j,p}(f)\right)\widehat{D}_{j,p-1}(\zeta).$$

Due to the linearity of the functionals $\lambda_{j,p}$, we have, for any given $\zeta^* \in \mathbb{R}$

$$\lambda_{j,p}(f) = \lambda_{j,p}\left(\int_0^{\zeta^*} g(s)\mathrm{d}s\right) + \lambda_{j,p}\left(\int_{\zeta^*}^{\zeta} g(s)\mathrm{d}s\right),$$

and noting that the term $\int_0^{\xi_j} g(s)\mathrm{d}s$ is a constant, thanks to the partition of unity of the B-spline functions $\widehat{B}_{i,p}$ it holds

$$\lambda_{j+1,p}\left(\int_0^{\xi_j} g(s)\mathrm{d}s\right) = \lambda_{j,p}\left(\int_0^{\xi_j} g(s)\mathrm{d}s\right).$$

Combining the last three equations, we obtain (2.23).

To prove (2.24), we use again the definition of $\Pi^c_{p-1,\Xi'}$, and then the stability of the projector $\Pi_{p,\Xi}$ from (2.15), to get

$$\|\Pi^c_{p-1,\Xi'} g\|_{L^2(I_i)} = |\Pi_{p,\Xi} f|_{H^1(I_i)} \le C|f|_{H^1(\widetilde{I}_i)} = C\|g\|_{L^2(\widetilde{I}_i)},$$

and the result is proved. □

Remark 2.5. *The same construction can be repeated replacing $\Pi_{p,\Xi}$ with $\widetilde{\Pi}_{p,\Xi}$ in the definition of $\Pi^c_{p-1,\Xi'}$, and we set:*

$$\widetilde{\Pi}^c_{p-1,\Xi'} g\,(\zeta) := \frac{d}{d\zeta}\widetilde{\Pi}_{p,\Xi} \int_0^\zeta g(s)\,\mathrm{d}s. \tag{2.25}$$

The operator $\widetilde{\Pi}^c_{p-1,\Xi'}$ enjoys the same properties as $\Pi^c_{p-1,\Xi'}$, i.e., the Proposition 2.4 holds verbatim also for $\widetilde{\Pi}^c_{p-1,\Xi'}$.

Remark 2.6. *Notice that the definition of the dual functional $\lambda^c_{j,p-1}$ depends on the local knot vectors $\Xi_{j,p}$ and $\Xi_{j+1,p}$, and therefore it goes beyond the support of $\widehat{D}_{j,p-1}$. Moreover, in the estimate* (2.24)*, the support extension $\widetilde{I}_i$ is defined for degree p, not $p-1$. This means that the projector $\Pi^c_{p-1,\Xi'}$ looses some locality with respect to $\Pi_{p-1,\Xi'}$, which would be the quasi-interpolant defined in [12, Section 4.6]. This is the "price to pay" in order to obtain the commuting diagram.*

2.3. Multivariate splines: tensorization. Multivariate B-splines are defined from univariate B-splines by simple tensorization. In this section, we basically set our notation for function spaces, basis functions and indices. Since tensorization argument is quite standard, we proceed without many details and we refer the reader to [12] and [13], or to the book [6].

Let d be the space dimension (in practical cases, $d = 2, 3$). Assume $n_\ell \in \mathbb{N}$, the degree $p_\ell \in \mathbb{N}$ and the p_ℓ-open knot vector $\Xi_\ell = \{\xi_{\ell,1}, \dots, \xi_{\ell,n_\ell+p_\ell+1}\}$ are given, for $\ell = 1, \dots, d$. We set the polynomial degree vector $\mathbf{p} = (p_1, \dots, p_d)$ and $\boldsymbol{\Xi} = \Xi_1 \times \ldots \times \Xi_d$. The corresponding knot values without repetitions are given for each direction ℓ by $Z_\ell = \{\zeta_{\ell,1}, \dots, \zeta_{\ell,N_\ell}\}$.

The knots Z_ℓ form a Cartesian grid in the *parametric domain* $\widehat{\Omega} = (0,1)^d$, giving the *parametric Bézier mesh*, which is denoted by $\widehat{\mathcal{M}}$:

$$\widehat{\mathcal{M}} = \{Q_{\mathbf{j}} = I_{1,j_1} \times \ldots \times I_{d,j_d} \text{ such that } I_{\ell,j_\ell} = (\zeta_{\ell,j_\ell}, \zeta_{\ell,j_\ell+1}) \text{ for } 1 \le j_\ell \le N_\ell - 1\}. \tag{2.26}$$

For a generic Bézier element $Q_{\mathbf{j}} \in \widehat{\mathcal{M}}$, we also define its *support extension* $\widetilde{Q}_{\mathbf{j}} = \widetilde{I}_{1,j_1} \times \ldots \times \widetilde{I}_{d,j_d}$, with $\widetilde{I}_{\ell,j_\ell}$ the univariate support extension given by (2.5). As in one space dimension, here also we make the following assumption:

Assumption 2.7. *Assumption 2.1 holds for each univariate partition , $j = 1, \dots, d$,*

B-spline spaces are defined by tensor product. We first introduce the set of multi-indices $\mathbf{I} = \{\mathbf{i} = (i_1, \dots, i_d) : 1 \le i_\ell \le n_\ell\}$, and for each multi-index $\mathbf{i} = (i_1, \dots, i_d)$, we introduce the of multivariate B-splines

$$S_{\mathbf{p}}(\boldsymbol{\Xi}) = S_{p_1,\dots,p_d}(\Xi_1, \dots, \Xi_d) = \operatorname{span}\{\widehat{B}_{i_1,p_1}(\zeta_1) \dots \widehat{B}_{i_d,p_d}(\zeta_d)\ \mathbf{i} \in \mathbf{I}\}. \tag{2.27}$$

The Greville sites, as in the univariate case, are the coefficients of the identity in the B-spline basis

$$\boldsymbol{\zeta} = \sum_{\mathbf{i}\in\mathbf{I}} \gamma_{\mathbf{i},\mathbf{p}} \widehat{B}_{\mathbf{i},\mathbf{p}}(\boldsymbol{\zeta}), \quad \boldsymbol{\zeta} \in \widehat{\Omega} = (0,1)^d. \tag{2.28}$$

and we denote by $\widehat{\mathcal{M}}_G$ the *Greville mesh* obtained by joining the Greville points in a tensor product mesh. Note that $\widehat{\mathcal{M}}_G$ is the tensorization of the Greville mesh defined in the previous section.

Finally, projection operators can be defined by tensorization, but this fact will be discussed later on.

3. The De Rham complex

Let Ω be a Lipschitz domain in $\mathbb{R}^3$, which we suppose for simplicity to be connected and simply connected. $L^2(\Omega)$ is the space of real valued, square integrable functions, and

$$\begin{aligned} \mathbf{H}(\mathbf{curl};\Omega) &:= \quad \{\mathbf{u} \in L^2(\Omega)^3 \ : \ \mathbf{curl}\,\mathbf{u} \in L^2(\Omega)^3\} \\ \mathbf{H}(\mathrm{div};\Omega) &:= \quad \{\mathbf{u} \in L^2(\Omega)^3 \ : \ \mathrm{div}\ \mathbf{u} \in L^2(\Omega)\}. \end{aligned}$$

On Ω, we define the spaces:

$$X^0 := H^1(\Omega)\,,\ X^1 := \mathbf{H}(\mathbf{curl};\Omega)\,,\ X^2 := \mathbf{H}(\mathrm{div};\Omega)\,,\ X^3 := L^2(\Omega)\,.$$

and it is well known that the sequence, known as De Rham diagram,

$$\mathbb{R} \longrightarrow X^0 \xrightarrow{\mathbf{grad}} X^1 \xrightarrow{\mathbf{curl}} X^2 \xrightarrow{\mathrm{div}} X^3 \longrightarrow 0. \tag{3.1}$$

is exact. When the domain Ω is indeed $\Omega = \widehat{\Omega} = (0,1)^d$, then we will denote the corresponding diagram as $(\widehat{X}^0, \ldots, \widehat{X}^3)$. Moreover, if there is a smooth $\mathbf{F} : \widehat{\Omega} \to \Omega$, with smooth inverse, then the pullbacks are defined (see [8, Sect. 2.2]):

$$\begin{array}{ll} \iota^0(f) := f \circ \mathbf{F}, & f \in X^0, \\ \iota^1(\mathbf{f}) := (D\mathbf{F})^T(\mathbf{f} \circ \mathbf{F}), & \mathbf{f} \in X^1, \\ \iota^2(\mathbf{f}) := \det(D\mathbf{F})(D\mathbf{F})^{-1}(\mathbf{f} \circ \mathbf{F}), & \mathbf{f} \in X^2, \\ \iota^3(f) := \det(D\mathbf{F})(f \circ \mathbf{F}), & f \in X^3, \end{array} \tag{3.2}$$

where $D\mathbf{F}$ is the Jacobian matrix of the mapping $\mathbf{F}$. Then, due to the curl and divergence conserving properties of ι^1 and ι^2, respectively (see [15, Sect. 3.9], for instance), the following commuting diagram commutes: (see [8, Sect. 2.2]):

$$\begin{array}{ccccccccc} \mathbb{R} & \longrightarrow & \widehat{X}^0 & \xrightarrow{\widehat{\mathbf{grad}}} & \widehat{X}^1 & \xrightarrow{\widehat{\mathbf{curl}}} & \widehat{X}^2 & \xrightarrow{\widehat{\mathrm{div}}} & \widehat{X}^3 & \longrightarrow & 0 \\ & & \iota^0 \uparrow & & \iota^1 \uparrow & & \iota^2 \uparrow & & \iota^3 \uparrow & & \\ \mathbb{R} & \longrightarrow & X^0 & \xrightarrow{\mathbf{grad}} & X^1 & \xrightarrow{\mathbf{curl}} & X^2 & \xrightarrow{\mathrm{div}} & X^3 & \longrightarrow & 0 \end{array} \tag{3.3}$$

where differential operators with a $\widehat{\cdot}$ stands for derivations in $\widehat{\Omega}$.

Remark 3.1. *There is an analogue of the sequence* (3.1) *involving spaces with boundary conditions on a part of the boundary $\Gamma_C \subset \partial\Omega$. All the theory and construction developed in this paper apply also to this case with minor changes.*

4. The Spline complex on the parametric domain

This section is devoted to the construction of the De Rham diagram in the unit cube $\widehat{\Omega} =]0,1[^3$.

First of all, using the expression for the derivative (2.7) in three dimensions, it is clear that, e.g.,

$$\frac{\partial}{\partial\zeta_1} : S_{p_1,p_2,p_3}(\Xi_1,\Xi_2,\Xi_3) \to S_{p_1-1,p_2,p_3}(\Xi_1',\Xi_2,\Xi_3)$$

where we remind that Ξ_1' is defined as the knot vector $\{\xi_{1,2},\dots,\xi_{1,n_1+p_1}\}$.

Following the same rationale, we define the spaces on the parametric domain $\widehat{\Omega}$:

$$\begin{aligned}
\widehat{X}_h^0 &:= S_{p_1,p_2,p_3}(\Xi_1,\Xi_2,\Xi_3),\\
\widehat{X}_h^1 &:= S_{p_1-1,p_2,p_3}(\Xi_1',\Xi_2,\Xi_3)\times S_{p_1,p_2-1,p_3}(\Xi_1,\Xi_2',\Xi_3)\times S_{p_1,p_2,p_3-1}(\Xi_1,\Xi_2,\Xi_3'),\\
\widehat{X}_h^2 &:= S_{p_1,p_2-1,p_3-1}(\Xi_1,\Xi_2',\Xi_3')\times S_{p_1-1,p_2,p_3-1}(\Xi_1',\Xi_2,\Xi_3')\\
&\quad\times S_{p_1-1,p_2-1,p_3}(\Xi_1',\Xi_2',\Xi_3),\\
\widehat{X}_h^3 &:= S_{p_1-1,p_2-1,p_3-1}(\Xi_1',\Xi_2',\Xi_3').
\end{aligned} \tag{4.1}$$

In order for $\widehat{X}_h^1$, $\widehat{X}_h^2$ and $\widehat{X}_h^3$ to be meaningful, we require $0 \le m_{\ell,i} \le p_\ell$, for $i = 2,\dots,N_\ell-1$ and $\ell = 1,2,3$. This means that the functions in $\widehat{X}_h^0$ are at least continuous. Then, thanks to (2.7) it is easily seen that $\widehat{\mathbf{grad}}\,(\widehat{X}_h^0) \subset \widehat{X}_h^1$, and analogously, from the definition of the curl and the divergence operators we get $\widehat{\mathbf{curl}}(\widehat{X}_h^1) \subset \widehat{X}_h^2$, and $\widehat{\mathrm{div}}\,(\widehat{X}_h^2) \subset \widehat{X}_h^3$. Moreover, it is proved in [2] that the kernel of each operator is exactly the image of the preceding one. In other words, these spaces form an exact sequence:

$$\mathbb{R} \longrightarrow \widehat{X}_h^0 \xrightarrow{\widehat{\mathbf{grad}}} \widehat{X}_h^1 \xrightarrow{\widehat{\mathbf{curl}}} \widehat{X}_h^2 \xrightarrow{\widehat{\mathrm{div}}} \widehat{X}_h^3 \longrightarrow 0, \tag{4.2}$$

that is, the first line of (3.3).

Remark 4.1. *The spaces defined above do not have boundary conditions, but all what we will present in this section can be extended with minor changes to spaces where homogeneous boundary conditions are applied to a set of faces $\widehat{\Gamma}_D$ of $\partial\widehat{\Omega}$.*

4.1. Choice of bases and topological structure. First of all, we define appropriate basis for the spline spaces. Our choice will make evident that the topological structure of the spline complex is closely related to the one of the Greville mesh $\widehat{\mathcal{M}}_G$ for the space $\widehat{X}_h^0$[1]. Let us start with a simple one-dimensional argument. Let us first remind the formula (2.8):

$$\frac{d\widehat{B}_{i,p}}{d\zeta}(\zeta) = \widehat{D}_{i-1,p-1}(\zeta) - \widehat{D}_{i,p-1}(\zeta).$$

This means that on the segment $[0,1]$, if we choose $\{\widehat{B}_{i,p}, i = 1,\dots,n\}$ and $\{\widehat{D}_{i,p-1}, i = 1,\dots,n-1\}$ as basis for $S_p(\Xi)$ and $S_{p-1}(\Xi')$, respectively, then the matrix representing the derivative is the lower triangular, bidiagonal matrix which represents the *vertex-to-edge* relation on the one-dimensional Greville mesh constructed from Ξ and p, i.e., with vertices given by (2.9). This means that we are implicitly setting an association between the edges of the Greville mesh and functions $\widehat{D}_{i,p-1}$.

As in [16] and [1], inspired by this observation, we can choose the following set of basis for the spaces in the spline complex:

$$\widehat{X}_h^0 = \text{span}\left\{\widehat{B}_{i_1,p_1}(\zeta_1)\widehat{B}_{i_2,p_2}(\zeta_2)\widehat{B}_{i_3,p_3}(\zeta_3) \text{ with } 1 \le i_\ell \le n_\ell,\ \ell = 1,2,3\right\}, \tag{4.3}$$

[1]Note that the Greville mesh $\widehat{\mathcal{M}}_G$ as defined in Section 2.3 is different for each space (and for each component)

$$
\begin{aligned}
\widehat{X}_h^1 &= \text{span}\,(I \cup II \cup III),\ \text{with}\\
I &= \Big\{ \widehat{D}_{i_1,p_1-1}(\zeta_1)\widehat{B}_{i_2,p_2}(\zeta_2)\widehat{B}_{i_3,p_3}(\zeta_3)\widehat{\mathbf{e}}_1 \text{ with}\\
&\qquad\qquad 1 \le i_1 \le n_1 - 1,\ 1 \le i_\ell \le n_\ell,\ \ell = 2,3\Big\},\\
II &= \Big\{ \widehat{B}_{i_1,p_1}(\zeta_1)\widehat{D}_{i_2,p_2-1}(\zeta_2)\widehat{B}_{i_3,p_3}(\zeta_3)\widehat{\mathbf{e}}_2 \text{ with}\\
&\qquad\qquad 1 \le i_2 \le n_2 - 1,\ 1 \le i_\ell \le n_\ell,\ \ell = 1,3\Big\},\\
III &= \Big\{ \widehat{B}_{i_1,p_1}(\zeta_1)\widehat{B}_{i_2,p_2}(\zeta_2)\widehat{D}_{i_3,p_3-1}(\zeta_3)\widehat{\mathbf{e}}_3 \text{ with}\\
&\qquad\qquad 1 \le i_3 \le n_3 - 1,\ 1 \le i_\ell \le n_\ell,\ \ell = 1,2\Big\},
\end{aligned}
\tag{4.4}
$$

$$
\begin{aligned}
\widehat{X}_h^2 &= \text{span}\,(I \cup II \cup III),\ \text{with}\\
I &= \Big\{ \widehat{B}_{i_1,p_1}(\zeta_1)\widehat{D}_{i_2,p_2-1}(\zeta_2)\widehat{D}_{i_3,p_3-1}(\zeta_3)\widehat{\mathbf{e}}_1 \text{ with}\\
&\qquad\qquad 1 \le i_1 \le n_1,\ 1 \le i_\ell \le n_\ell - 1,\ \ell = 2,3\Big\},\\
II &= \Big\{ \widehat{D}_{i_1,p_1-1}(\zeta_1)\widehat{B}_{i_2,p_2}(\zeta_2)\widehat{D}_{i_3,p_3-1}(\zeta_3)\widehat{\mathbf{e}}_2 \text{ with}\\
&\qquad\qquad 1 \le i_2 \le n_2,\ 1 \le i_\ell \le n_\ell - 1,\ \ell = 1,3\Big\},\\
III &= \Big\{ \widehat{D}_{i_1,p_1-1}(\zeta_1)\widehat{D}_{i_2,p_2-1}(\zeta_2)\widehat{B}_{i_3,p_3}(\zeta_3)\widehat{\mathbf{e}}_3 \text{ with}\\
&\qquad\qquad 1 \le i_3 \le n_3,\ 1 \le i_\ell \le n_\ell - 1,\ \ell = 1,2\Big\},
\end{aligned}
\tag{4.5}
$$

$$
\widehat{X}_h^3 = \text{span}\Big\{ \widehat{D}_{i_1,p_1-1}(\zeta_1)\widehat{D}_{i_2,p_2-1}(\zeta_2)\widehat{D}_{i_3,p_3-1}(\zeta_3) \text{ with } 1 \le i_\ell \le n_\ell - 1,\ \ell = 1,2,3\Big\},
\tag{4.6}
$$

where $\{\widehat{\mathbf{e}}_\ell\}_{\ell=1,2,3}$ denote the canonical basis of $\mathbb{R}^3$. We remark that all basis functions defined in (4.3)-(4.6) are non-negative.

Moreover, by using the formula (2.8) and a tensor product argument, we can analyse the structure of the basis functions. For instance, we consider the set I in (4.4). By construction, we have one of these functions per each edge of $\widehat{\mathcal{M}}_G$ in the ζ_1-direction, and these functions are directed as the edges. Applying the same reasoning to the other set of basis functions in (4.3-4.6), together with the structure of the matrices representing differential operators, we have the following:

Proposition 4.2. *With the choices (4.3-4.6), the matrices representing differential operators* $\widehat{\mathbf{grad}}$, $\widehat{\mathbf{curl}}$, *and* $\widehat{\mathrm{div}}$ *are the incidence matrices of the tensor product mesh* $\widehat{\mathcal{M}}_G$. *Thus, the spline complex* $(\widehat{X}_h^0, \widehat{X}_h^1, \widehat{X}_h^2, \widehat{X}_h^3)$ *is isomorphic to the co-chain complex associated with mesh* $\widehat{\mathcal{M}}_G$.

The previous proposition states that the spline complex has exactly the same structure of the well known Whitney forms when defined on the tensor product mesh $\widehat{\mathcal{M}}_G$, see, e.g. Section 3 in [8].

Remark 4.3. *Let* $p = p_1 = p_2 = p_3$. *If all knots are repeated* p *times, the formulae (4.4-4.6) provide a canonical construction of basis for the standard finite element complex of order* p, *see e.g., [9] on the partition* $\widehat{\mathcal{M}}$ *counted without its repetition.*

Remark 4.4. *It is also possible to study the relation between the complex* $(\widehat{X}_h^0, \dots, \widehat{X}_h^3)$ *and the topological structure of the mesh* $\widehat{\mathcal{M}}$*. This is done in the paper [1] and the following interesting fact is true:*

- *when* p *is odd, the complex* $(\widehat{X}_h^0, \dots, \widehat{X}_h^3)$ *is isomorphic to the co-chain complex of* $\widehat{\mathcal{M}}$*, counted with its repetition,*
- *when* p *is even, the complex* $(\widehat{X}_h^0, \dots, \widehat{X}_h^3)$ *is isomorphic to the chain complex of* $\widehat{\mathcal{M}}$*, counted with its repetition. This fact has in principle a number of applications, as, for example, to the preconditioning of integral equations [17].*

4.2. Commuting projections. It is now necessary to define appropriate projectors into the discrete spaces. This is done by using the definition of interpolants and quasi-interpolants that we have given in Section 2.2. To alleviate notation, from this point we will not detail the knot vector in the interpolant, that is, we will denote $\Pi_p \equiv \Pi_{p,\Xi}$ and $\Pi^c_{p-1} \equiv \Pi^c_{p-1,\Xi'}$. The choice of the interpolants follows from the definition of the spaces $\widehat{X}_h^0$, …, $\widehat{X}_h^3$, and precisely we set:

$$\widehat{\Pi}^0 := \Pi_{p_1} \otimes \Pi_{p_2} \otimes \Pi_{p_3}, \tag{4.7}$$

$$\widehat{\Pi}^1 := (\Pi^c_{p_1-1} \otimes \Pi_{p_2} \otimes \Pi_{p_3}) \times (\Pi_{p_1} \otimes \Pi^c_{p_2-1} \otimes \Pi_{p_3}) \times (\Pi_{p_1} \otimes \Pi_{p_2} \otimes \Pi^c_{p_3-1}), \tag{4.8}$$

$$\begin{aligned}\widehat{\Pi}^2 := &(\Pi_{p_1} \otimes \Pi^c_{p_2-1} \otimes \Pi^c_{p_3-1}) \times \\ &(\Pi^c_{p_1-1} \otimes \Pi_{p_2} \otimes \Pi^c_{p_3-1}) \times (\Pi^c_{p_1-1} \otimes \Pi^c_{p_2-1} \otimes \Pi_{p_3}),\end{aligned} \tag{4.9}$$

$$\widehat{\Pi}^3 := \Pi^c_{p_1-1} \otimes \Pi^c_{p_2-1} \otimes \Pi^c_{p_3-1}. \tag{4.10}$$

Remark 4.5. *It should be noted that if we replace* Π_{p_ℓ} *with* $\widetilde{\Pi}_{p_\ell}$ *and* $\Pi^c_{p_\ell-1}$ *with* $\widetilde{\Pi}^c_{p-1}$*, then we define another set of projectors that enjoys all the properties described here below. Moreover, this choice will be useful to define projectors in some special case later on.*

The next lemma shows that the interpolants are projectors onto the corresponding spline spaces.

Lemma 4.6. *The interpolants* (4.7)-(4.10) *satisfy the spline preserving property, that is*

$$\begin{aligned}\widehat{\Pi}^i \widehat{f}_h = \widehat{f}_h, \quad \forall \widehat{f}_h \in \widehat{X}^i_h,\ i = 0, 3,\\ \widehat{\Pi}^i \widehat{\mathbf{f}}_h = \widehat{\mathbf{f}}_h, \quad \forall \widehat{\mathbf{f}}_h \in \widehat{X}^i_h,\ i = 1, 2.\end{aligned}$$

Proof. The result is an immediate consequence of the splines preserving property of the interpolants Π_{p_ℓ} and $\Pi^c_{p_\ell-1}$, $\ell = 1, 2, 3$, given in (2.13) and in (2.21), respectively. □

Lemma 4.7. *Under Assumption 2.7, the following inequalities hold for any* $Q \in \widehat{\mathcal{M}}$*:*

$$\begin{aligned}\|\widehat{\Pi}^i \widehat{f}\|_{L^2(Q)} \le C \|\widehat{f}\|_{L^2(\widetilde{Q})} \quad \forall \widehat{f} \in L^2(\widehat{\Omega}), \quad i = 0, 3,\\ \|\widehat{\Pi}^i \widehat{\mathbf{f}}\|_{L^2(Q)^3} \le C \|\widehat{\mathbf{f}}\|_{L^2(\widetilde{Q})^3} \quad \forall \widehat{\mathbf{f}} \in L^2(\widehat{\Omega})^3, \quad i = 1, 2.\end{aligned}$$

Proof. The result follows immediately from (2.14) and (2.24), which state that the involved one-dimensional operators Π_{p_ℓ} and $\Pi^c_{p_\ell-1}$, $\ell = 1, 2, 3$ are L^2 stable. □

The commutativity of the interpolants with the differential operators is stated in the following lemma.

Lemma 4.8. *It holds*

$$\widehat{\mathbf{grad}}\,(\widehat{\Pi}^0\widehat{f}) = \widehat{\Pi}^1(\widehat{\mathbf{grad}}\,\widehat{f}) \qquad \forall \widehat{f} \in \widehat{X}^0, \tag{4.11}$$

$$\widehat{\mathbf{curl}}(\widehat{\Pi}^1\widehat{\mathbf{f}}) = \widehat{\Pi}^2(\widehat{\mathbf{curl}}\,\widehat{\mathbf{f}}) \qquad \forall \widehat{\mathbf{f}} \in \widehat{X}^1, \tag{4.12}$$

$$\widehat{\mathrm{div}}\,(\widehat{\Pi}^2\widehat{\mathbf{f}}) = \widehat{\Pi}^3(\widehat{\mathrm{div}}\,\widehat{\mathbf{f}}) \qquad \forall \widehat{\mathbf{f}} \in \widehat{X}^2. \tag{4.13}$$

Proof. The proof is based on the commutativity property (2.20) and the tensor product structure of the spaces and interpolants. Consider first (4.11): let $\widehat{f}$ be a smooth scalar field with compact support in $\widehat{\Omega}$. The first component of $\widehat{\mathbf{grad}}\,(\widehat{\Pi}^0\widehat{f})$ is given by

$$\begin{aligned}\partial_{\widehat{x}}(\widehat{\Pi}^0\widehat{f}) &= \partial_{\widehat{x}}((\Pi_{p_1}\otimes\Pi_{p_2}\otimes\Pi_{p_3})\,\widehat{f}) = \partial_{\widehat{x}}(\Pi_{p_1}(\Pi_{p_2}(\Pi_{p_3}\,\widehat{f})))\\ &= \Pi^c_{p_1-1}\partial_{\widehat{x}}(\Pi_{p_2}(\Pi_{p_3}\,\widehat{f})) = (\Pi^c_{p_1-1}\otimes\Pi_{p_2}\otimes\Pi_{p_3})\,\partial_{\widehat{x}}\widehat{f},\end{aligned}$$

which is the first component of $\widehat{\Pi}^1(\widehat{\mathbf{grad}}\,\widehat{f})$. A similar reasoning, using the commutativity of the univariate interpolants, yields

$$\begin{aligned}\partial_{\widehat{y}}(\widehat{\Pi}^0\widehat{f}) &= (\Pi_{p_1}\otimes\Pi^c_{p_2-1}\otimes\Pi_{p_3})\,\partial_{\widehat{y}}\widehat{f},\\ \partial_{\widehat{z}}(\widehat{\Pi}^0\widehat{f}) &= (\Pi_{p_1}\otimes\Pi_{p_2}\otimes\Pi^c_{p_3-1})\,\partial_{\widehat{z}}\widehat{f},\end{aligned}$$

which proves that $\widehat{\mathbf{grad}}\,(\widehat{\Pi}^0\widehat{f}) = \widehat{\Pi}^1(\widehat{\mathbf{grad}}\,\widehat{f})$. By a density argument (4.11) follows, thanks to Lemma 4.7. The proof of (4.12)–(4.13) is similar, from the definition of the interpolants and the expression of the curl and divergence operators. □

4.3. Approximation estimates. This section is devoted to the study of the approximation estimates of the complex $(\widehat{X}^0_h,\dots,\widehat{X}^3_h)$. The content of this section is based on the paper [2].

We start from the definition of the bent Sobolev spaces that we need. Since the interelement regularity changes from space to space (and from component to component), we need here to make the notation more explicit, starting from the one-dimensional definition: we denote by $\mathcal{H}^s_{\mathbf{k}}(I)$, $I=(0,1)$, the space defined as:

$$\mathcal{H}^s_{\mathbf{k}}(I) = \left\{ \begin{array}{l} f\in L^2(I) \text{ such that } f|_{I_i}\in H^s(I_i)\;\forall\, i=1,\dots,N-1, \text{ and}\\ D^k_- f(\zeta_i) = D^k_+ f(\zeta_i),\ \forall k=0,\dots,\min\{s-1,k_i\}, \forall i=2,\dots,N-1, \end{array}\right\} \tag{4.14}$$

where $\mathbf{k}=(k_2,\dots,k_{N-1})$ and k_i is the number of continuous derivatives at the point $\zeta_i\in Z$.

In three dimensions, given $\mathbf{s}\in\mathbb{N}^3$ and the three vectors $\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3$ constructed from $\boldsymbol{\Xi}$, we set:

$$\mathcal{H}^{\mathbf{s}}_{\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3} = \mathcal{H}^{s_1}_{\mathbf{k}_1}(I)\otimes\mathcal{H}^{s_2}_{\mathbf{k}_2}(I)\otimes\mathcal{H}^{s_3}_{\mathbf{k}_3}(I),$$

and also:

$$\begin{aligned}\mathcal{H}^{0,\mathbf{s}} &= \mathcal{H}^{\mathbf{s}}_{\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3}\\ \mathcal{H}^{1,\mathbf{s}} &= \mathcal{H}^{\mathbf{s}}_{\mathbf{k}_1-1,\mathbf{k}_2,\mathbf{k}_3}\times\mathcal{H}^{\mathbf{s}}_{\mathbf{k}_1,\mathbf{k}_2-1,\mathbf{k}_3}\times\mathcal{H}^{\mathbf{s}}_{\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3-1}\\ \mathcal{H}^{2,\mathbf{s}} &= \mathcal{H}^{\mathbf{s}}_{\mathbf{k}_1,\mathbf{k}_2-1,\mathbf{k}_3-1}\times\mathcal{H}^{\mathbf{s}}_{\mathbf{k}_1-1,\mathbf{k}_2,\mathbf{k}_3-1}\times\mathcal{H}^{\mathbf{s}}_{\mathbf{k}-1,\mathbf{k}_2-1,\mathbf{k}_3}\\ \mathcal{H}^{3,\mathbf{s}} &= \mathcal{H}^{\mathbf{s}}_{\mathbf{k}-1,\mathbf{k}_2-1,\mathbf{k}_3-1}.\end{aligned} \tag{4.15}$$

This choice is made in order to ensure that $\widehat{X}_h^i \subset \mathcal{H}^{i,\mathbf{s}}$, for all $\mathbf{s} \in \mathbb{N}^3$, $i = 1, 2, 3$, i.e., the interelement regularity of $\mathcal{H}^{i,\mathbf{s}}$ is *not higher* than the one of $\widehat{X}_h^i$. The semi-norm corresponding to $\mathcal{H}^{0,\mathbf{s}}$ is defined as $|f|^2_{\mathcal{H}^{0,\mathbf{s}}(D)} = \sum_{Q\in\widehat{\mathcal{M}}\cap D} |f|^2_{H^{\mathbf{s}}(Q)}$. Moreover, we define

$$|f|^2_{\mathcal{H}^{0,|\mathbf{s}|}(\widehat{\Omega})} = \sup_{\mathbf{r}\,:\,|\mathbf{r}|\leq|\mathbf{s}|} |f|^2_{\mathcal{H}^{0,\mathbf{s}}(\widehat{\Omega})}; \tag{4.16}$$

while the norms for $\mathcal{H}^{1,\mathbf{s}}$ and $\mathcal{H}^{2,\mathbf{s}}$, and the corresponding norms $\|\cdot\|_{\mathcal{H}^{1,|\mathbf{s}|}}$ and $\|\cdot\|_{\mathcal{H}^{2,|\mathbf{s}|}}$, are defined component by component in a similar way. For simplicity, we write estimates that depends only on $p = \min\{p_1, p_2, p_3\}$ and $|\mathbf{s}| = s_1 + s_2 + s_3$. The following holds:

Proposition 4.9. *Let Assumption 2.7 hold, Q be an element of $\widehat{\mathcal{M}}$, and $\widetilde{Q}$ its extension. Then it holds, for $i = 0, 3$,*

$$\|(\widehat{f} - \widehat{\Pi}^i \widehat{f})\|_{H^r(Q)^3} \leq C h_{\widetilde{Q}}^{|\mathbf{s}|-r} \, |\widehat{f}|_{\mathcal{H}^{i,|\mathbf{s}|}(\widetilde{Q})} \tag{4.17}$$

for all $\widehat{f} \in \mathcal{H}^{i,\mathbf{t}}$, for all $\mathbf{t}$, $|\mathbf{t}| \leq |\mathbf{s}|$, and when $i = 0$, $0 \leq r \leq |\mathbf{s}| \leq p+1$, while when $i = 3$, $0 \leq r \leq |\mathbf{s}| \leq p$.

And for $i = 1, 2$ it holds

$$\|(\widehat{\mathbf{f}} - \widehat{\Pi}^i \widehat{\mathbf{f}})\|_{H^r(Q)^3} \leq C h_{\widetilde{Q}}^{|\mathbf{s}|-r} \, |\widehat{\mathbf{f}}|_{\mathcal{H}^{i,|\mathbf{s}|}(\widetilde{Q})} \tag{4.18}$$

for all $\widehat{\mathbf{f}} \in \mathcal{H}^{i,\mathbf{t}}$ for all $\mathbf{t}$, $|\mathbf{t}| \leq |\mathbf{s}|$ and $0 \leq r \leq |\mathbf{s}| \leq p$.

Proof. The proof of this statement can be found in e.g., [4] where indeed a more general estimate is proposed. □

Remark 4.10. *A similar result holds also if we replace Π_{p_ℓ} with $\widetilde{\Pi}_{p_\ell}$ and $\Pi^c_{p_\ell-1}$ with $\widetilde{\Pi}^c_{p_\ell-1}$ (see Remark 4.5), with the only difference that the constraints on the allowed Sobolev index are more restrictive. In particular, for $i = 0$,* (4.17) *holds for $s > 3/2$ and for $i = 2, 3$,* (4.18) *holds only for $s > 1$.*

5. The spline complex on general domains

We suppose that we are given a domain Ω which is an open, bounded and connected and simply connected set, and which is defined as the union of M_p subdomains, in the form

$$\overline{\Omega} = \bigcup_{j=1}^{M_p} \overline{\Omega^{(j)}}, \tag{5.1}$$

where the subdomains $\Omega^{(j)} = \mathbf{F}^{(j)}(\widehat{\Omega})$ are referred to as *patches*, and are assumed to be disjoint. Each patch is obtained as a spline mapping of the reference domain $\widehat{\Omega}$. I.e., there is parametrization $\mathbf{F}^{(j)} : \widehat{\Omega} \to \Omega^{(j)}$ defined using a spline space $S_{p^j}(\boldsymbol{\Xi}^{(j)})$, on the parametric mesh $\widehat{\mathcal{M}}^{(j)}$. Remark that these mappings could be chosen as NURBS (see [7]) without any change in what follows. We note that

$$\mathbf{F}^{(j)} = \sum_{\mathbf{k}} \mathbf{c}_{\mathbf{k}}^{(j)} \widehat{B}^{(j)}_{\mathbf{k},\mathbf{p}}(\boldsymbol{\zeta})$$

where $\mathbf{c}_{\mathbf{k}}^{(j)}$ are the control points of $\mathbf{F}^{(j)}$ and the basis functions $\widehat{B}_{\mathbf{k},\mathbf{p}}^{(j)}(\boldsymbol{\zeta})$ depends on the choice of the knot vector $\boldsymbol{\Xi}^{(j)}$.

In what follows we suppose that each $\mathbf{F}^{(j)}$ verifies the following assumption:

Assumption 5.1 (Regularity of $\mathbf{F}^{(j)}$). *The parametrization $\mathbf{F}^{(j)} : \widehat{\Omega} \to \Omega^{(j)}$ is a bi-Lipschitz homeomorphism. Moreover, $\mathbf{F}^{(j)}_{|\bar{Q}}$ is in $C^\infty(\bar{Q})$ for all $Q \in \widehat{\mathcal{M}}^{(j)}$, where $\bar{Q}$ denotes the closure of Q, and $(\mathbf{F}^{(j)})^{-1}_{|\bar{K}}$ is in $C^\infty(\bar{K})$ for all $\bar{K} = (\mathbf{F}^{(j)})^{-1}(\bar{Q})$, $Q \in \widehat{\mathcal{M}}^{(j)}$.*

In each patch $\Omega^{(j)}$, there is a natural mesh, called *Bézier mesh*, as the image of the (open) elements in $\widehat{\mathcal{M}}^{(j)}$ through $\mathbf{F}^{(j)}$:

$$\mathcal{M}^{(j)} := \{K \subset \Omega : K = \mathbf{F}^{(j)}(Q), Q \in \widehat{\mathcal{M}}^{(j)}\}, \tag{5.2}$$

For any element $K = \mathbf{F}^{(j)}(Q) \in \mathcal{M}^{(j)}$, we define its support extension as $\widetilde{K} = \mathbf{F}(\widetilde{Q})$, with $\widetilde{Q}$ the support extension of Q, defined in Section 2.3. Moreover, we denote the element size of any element $Q \in \widehat{\mathcal{M}}^{(j)}$ by $h_Q = \operatorname{diam}(Q)$, and the global mesh size is $h = \max\{h_Q : Q \in \widehat{\mathcal{M}}\}$. Analogously, we define the element sizes $h_K = \operatorname{diam}(K)$ and $h_{\widetilde{K}} = \operatorname{diam}(\widetilde{K})$. Assumption 5.1 below will ensure that $h_Q \simeq h_K$.

Moreover, each scalar basis function $\widehat{B}_{\mathbf{k},\mathbf{p}}^{(j)}(\boldsymbol{\zeta})$ is mapped on $\Omega^{(j)}$ by simple change of variable:

$$\widehat{B}_{\mathbf{k},\mathbf{p}}^{(j)}(\boldsymbol{\zeta}) = B_{\mathbf{k},\mathbf{p}}^{(j)}(\mathbf{x}), \qquad \mathbf{x} = \mathbf{F}^{(j)}(\boldsymbol{\zeta}).$$

Finally, we call *control mesh*, the structured mesh $\mathcal{M}_C^{(j)}$ obtained by joining the control point $\mathbf{c}_{\mathbf{k}}^{(j)}$. Note that this mesh has exactly the same structure of $\widehat{\mathcal{M}}_G$, and we defer the reader [4] for a discussion on this.

In order to guarantee conformity in the construction of spline spaces on Ω, we have the following assumption:

Assumption 5.2 (Conformity). *Let $\Gamma_{ij} = \partial\Omega^{(i)} \cap \partial\Omega^{(j)}$ be the interface between the patches $\Omega^{(i)}$ and $\Omega^{(j)}$, with $i \neq j$. We say that the two patches are* fully matching *if the two following conditions hold.*

(i) *Γ_{ij} is either a vertex, or the image of a full edge, or the image of a full face for both parametric domains.*

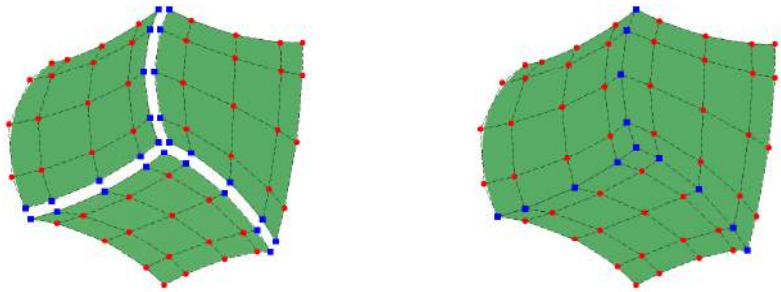
(ii) *For each basis functions $B_{\mathbf{k},\mathbf{p}}^{(i)}$ such that $\operatorname{supp}(B_{\mathbf{k},\mathbf{p}}^{(i)}) \cap \Gamma_{ij} \neq \emptyset$, there exists a basis function $B_{\mathbf{l},\mathbf{p}}^{(j)}$ such that $B_{\mathbf{k},\mathbf{p}}^{(i)}|_{\Gamma_{ij}} = B_{\mathbf{l},\mathbf{p}}^{(j)}|_{\Gamma_{ij}}$ (and viceversa). Moreover, the related control points $\mathbf{c}_{\mathbf{k}}^{(i)}$ and $\mathbf{c}_{\mathbf{l}}^{(j)}$ coincide: $\mathbf{c}_{\mathbf{k}}^{(i)} = \mathbf{c}_{\mathbf{l}}^{(j)}$.*

First of all, Assumption 5.2 means that the control mesh $\mathcal{M}_C = \cup_{j=1}^{M_p} \mathcal{M}_C^{(j)}$ is a conforming mesh.

Moreover, the physical Bézier meshes $\mathcal{M}^{(i)}$ and $\mathcal{M}^{(j)}$ coincide on the interface Γ_{ij}, and the coincident knot vectors are affinely related, including knot repetitions. Thus the partition

$$\mathcal{M} = \cup_{j=1}^{M_p} \mathcal{M}^{(j)} \tag{5.3}$$

is a *conforming*, globally unstructured, locally (to each patch) structured mesh of the computational domain Ω.



(a) Control mesh of the three separate patches (b) Control mesh of the multi-patch domain

Figure 5.1. Generation of a multi-patch domain with conforming meshes. The square control points are associated to basis functions that match on the interface.

We focus now on the definition of spline spaces that form a complex. For each patch $\Omega^{(j)}$, we set:

$$\begin{aligned} X_h^{0,(j)} &:= \{\phi_h : \iota^0(\phi_h) \in \widehat{X}_h^{0,(j)}\}, \\ X_h^{1,(j)} &:= \{\mathbf{u}_h : \iota^1(\mathbf{u}_h) \in \widehat{X}_h^{1,(j)}\}, \\ X_h^{2,(j)} &:= \{\mathbf{v}_h : \iota^2(\mathbf{v}_h) \in \widehat{X}_h^{2,(j)}\}, \\ X_h^{3,(j)} &:= \{\psi_h : \iota^3(\psi_h) \in \widehat{X}_h^{3,(j)}\}. \end{aligned} \tag{5.4}$$

where the spaces on the parametric domain are indexed with (j) because they depend on the parametric mesh $\widehat{\mathcal{M}}^{(j)}$, but are the ones constructed in (4.1). As a consequence of our choice, the spaces $(X_h^{0,(j)}, \dots, X_h^{3,(j)})$ form a complex and that the following holds:

$$\begin{array}{ccccccccc} \mathbb{R} & \longrightarrow & \widehat{X}_h^{0,(j)} & \xrightarrow{\widehat{\mathbf{grad}}} & \widehat{X}_h^{1,(j)} & \xrightarrow{\widehat{\mathbf{curl}}} & \widehat{X}_h^{2,(j)} & \xrightarrow{\widehat{\mathrm{div}}} & \widehat{X}_h^{3,(j)} & \longrightarrow & 0 \\ & & \iota^0 \uparrow & & \iota^1 \uparrow & & \iota^2 \uparrow & & \iota^3 \uparrow & & \\ \mathbb{R} & \longrightarrow & X_h^{0,(j)} & \xrightarrow{\mathbf{grad}} & X_h^{1,(j)} & \xrightarrow{\mathbf{curl}} & X_h^{2,(j)} & \xrightarrow{\mathrm{div}} & X_h^{3,(j)} & \longrightarrow & 0. \end{array} \tag{5.5}$$

On the domain Ω, we naturally construct:

$$\begin{aligned} X_h^0(\Omega) &= \{f \in H^1(\Omega) \ : \ f_{|\Omega^{(j)}} \in X_h^{0,(j)}\} \\ X_h^1(\Omega) &= \{\mathbf{f} \in \mathbf{H}(\mathbf{curl}; \Omega) \ : \ \mathbf{f}_{|\Omega^{(j)}} \in X_h^{1,(j)}\} \\ X_h^2(\Omega) &= \{\mathbf{f} \in \mathbf{H}(\mathrm{div}; \Omega) \ : \ \mathbf{f}_{|\Omega^{(j)}} \in X_h^{2,(j)}\} \\ X_h^3(\Omega) &= \{f \in L^2(\Omega) \ : \ f_{|\Omega^{(j)}} \in X_h^{3,(j)}\}. \end{aligned} \tag{5.6}$$

Having conformity of the control mesh $\mathcal{M}_C$, the continuity condition is implemented very easily by generating a global numbering, in a process that resembles the generation of the connectivity array in finite element meshes. For each non-empty interface Γ_{ij}, we

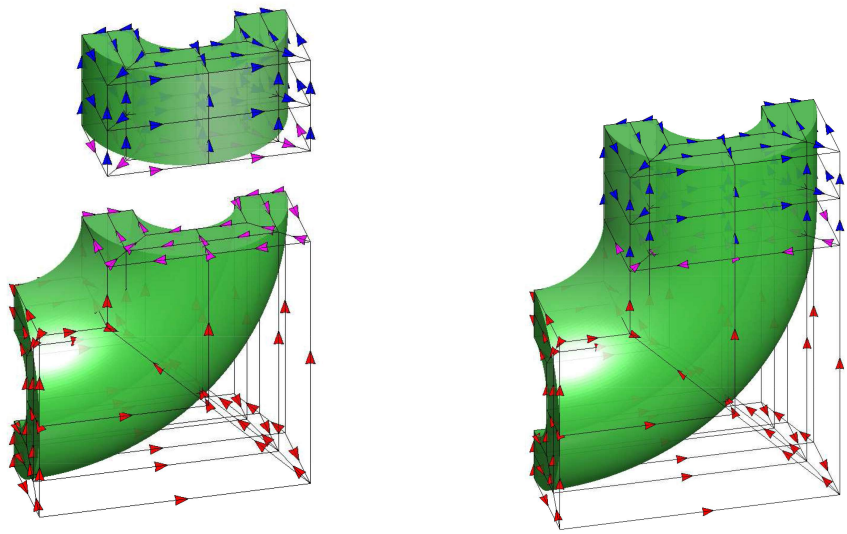
collect the pairs of coincident basis functions $B^{(i)}_{\mathbf{k},\mathbf{p}}$ and $B^{(j)}_{\mathbf{l},\mathbf{p}}$, and identify them as one single function, constraining their associated degrees of freedom to coincide. Note that for corners and edges (in the three-dimensional case), the new function may be generated from the contribution of functions coming from more than two patches.

More rigorously, we define for each patch $\Omega^{(j)}$, and precisely for the multi-index set $\mathbf{I}^{(j)}$, an application $G^{(j)} : \mathbf{I}^{(j)} \to \mathcal{I} = \{1, \dots, N_\Omega\}$, in such a way that $G^{(i)}(\mathbf{k}) = G^{(j)}(\mathbf{l}) \Leftrightarrow \Gamma_{ij} \neq \emptyset$ and $B^{(i)}_{\mathbf{k},\mathbf{p}}|_{\Gamma_{ij}} = B^{(j)}_{\mathbf{l},\mathbf{p}}|_{\Gamma_{ij}}$. The scalar N_Ω is the dimension of $X^0_h(\Omega)$, which is equal to the number of vertices of the control mesh $\mathcal{M}_C$. Moreover, we define for each global index $\ell \in \mathcal{I}$ the set of pairs $\mathcal{I}_\ell = \{(j, \mathbf{k}) : G^{(j)}(\mathbf{k}) = \ell\}$, which collects the local contributions to the global function. To conclude we define, for each $\ell \in \mathcal{I}$, the global basis function

$$B_\ell(\mathbf{x}) := \begin{cases} B^{(j)}_{\mathbf{k},\mathbf{p}}(\mathbf{x}) & \text{if } \mathbf{x} \in \overline{\Omega^{(j)}} \text{ and } (j, \mathbf{k}) \in \mathcal{I}_\ell, \\ 0 & \text{otherwise,} \end{cases} \tag{5.7}$$

which is continuous due to Assumption 5.2, and it holds that $X^0_h(\Omega) = \operatorname{span}\{B_\ell(\mathbf{x}) : \ell \in \mathcal{I}\}$.

Similarly, vector fields in $X^1_h(\Omega)$ (or $X^2_h(\Omega)$) are obtained by identifying the control variables associated to edges (or faces, respectively) that have been identified in the construction of $\mathcal{M}_C$, with a possible change of the orientation. In Figure 5.2, we describe this identification for the space $X^1_h(\Omega)$. Finally, functions in $X^3_h(\Omega)$ are discontinuous across the patch interfaces.



(a) Control variables of the two patches.

(b) Control variables of the merged patch.

Figure 5.2. Implementing continuity for $X^1_h(\Omega)$ on a two-patch domain. The orientation of the degrees of freedom associated to the interface edges (purple arrows) is chosen as that of the lower patch. The orientation for degrees of freedom not on the interface (blue and red arrows) remains unchanged after merging.

This construction is exactly the one that would be performed in the finite element context, but we remind that this standard procedure when applied to the control mesh $\mathcal{M}_C$ provides the correct interface condition for the underlying spline complex.

It is very easy to see that the following holds:

Proposition 5.3. *Under Assumptions 5.1 and 5.2, the spaces* $(X_h^0(\Omega), \dots, X_h^3(\Omega))$ *form a complex:*

$$\mathbb{R} \to X_h^0(\Omega) \xrightarrow{\mathbf{grad}} X_h^1(\Omega) \xrightarrow{\mathbf{curl}} X_h^2(\Omega) \xrightarrow{\mathrm{div}} X_h^3(\Omega) \to 0. \tag{5.8}$$

We end the section with the following important remark: thanks to (5.5), to Proposition 4.2, to the choice of basis functions in the parametric space, and to the control point identification, the differential operators in (5.8) are the incidence matrices of the control mesh $\mathcal{M}_C$.

This means, for example, that the spline complex is isomorphic to the complex of whitney forms (or low degree finite elements) on the mesh $\mathcal{M}_C$, as are defined in [9], but (and this is an important but) they deliver approximation rates of order p while the corresponding finite element complex may provide only linear convergence.

Remark 5.4. *Finally, we remark that all what results presented in the previous sections can be generalized to domains* Ω *that have a non trivial topology, stating that the spline sequence is a subcomplex of the De Rham complex and thus has the same topological structure.*

6. Approximation estimates on the physical domain

The first step is the study of the approximation estimate for a single patch domain and for the time being we drop the superindex $^{(j)}$, and set $M_p = 1$. We start by introducing the projectors for each space X_h^i of the complex. These projectors are defined from the ones in the parametric domain (4.7 - 4.10), and the corresponding pull-backs ι^i, in such a way that they are uniquely characterized by the equations

$$\begin{aligned} \iota^i(\Pi^i f) &= \widehat{\Pi}^i(\iota^i(f)) \quad i = 0, 3, \\ \iota^i(\Pi^i \mathbf{f}) &= \widehat{\Pi}^i(\iota^i(\mathbf{f})) \quad i = 1, 2. \end{aligned} \tag{6.1}$$

The following proposition is an immediate consequence of these definitions, together with the commuting properties of Lemma 4.8.

Proposition 6.1. *The following diagram commutes.*

$$\begin{array}{ccccccccccc} \mathbb{R} & \longrightarrow & X^0 & \xrightarrow{\mathbf{grad}} & X^1 & \xrightarrow{\mathbf{curl}} & X^2 & \xrightarrow{\mathrm{div}} & X^3 & \longrightarrow & 0 \\ & & \Pi^0 \downarrow & & \Pi^1 \downarrow & & \Pi^2 \downarrow & & \Pi^3 \downarrow & & \\ \mathbb{R} & \longrightarrow & X_h^0 & \xrightarrow{\mathbf{grad}} & X_h^1 & \xrightarrow{\mathbf{curl}} & X_h^2 & \xrightarrow{\mathrm{div}} & X_h^3 & \longrightarrow & 0. \end{array} \tag{6.2}$$

We first prove the following proposition, see also Lemma 3.6 in [2].

Proposition 6.2. *Let Assumption 2.7 and 5.1 hold. Let* $s \in \mathbb{N}$ *and* $\mathbf{s} = (s_1, s_2, s_3)$ *be any vector such that* $|\mathbf{s}| = s$*, and* $f \in H^s(\Omega)$*, and* $\mathbf{f} \in H^s(\Omega)^3$*. Then*

$$\begin{aligned} \iota^i(f) &\in \mathcal{H}^{i,\mathbf{s}} \qquad i = 0, 3, \\ \iota^i(\mathbf{f}) &\in \mathcal{H}^{i,\mathbf{s}} \qquad i = 1, 2. \end{aligned} \tag{6.3}$$

Moreover, there exists a constant C such that for all elements $K = \mathbf{F}(Q) \in \mathcal{M}$, with $Q \in \widehat{\mathcal{M}}$, it holds:

$$C^{-1}\|f\|_{H^s(K)} \le \|\iota^i(f)\|_{H^s(Q)} \le C\|f\|_{H^s(K)} \qquad i = 0, 3,$$
$$C^{-1}\|\mathbf{f}\|_{H^s(K)^3} \le \|\iota^i(\mathbf{f})\|_{H^s(Q)^3} \le C\|\mathbf{f}\|_{H^s(K)^3} \qquad i = 1, 2.$$

Proof. First of all, we show (6.3), and we focus on the case $i = 1$ since all other cases are similar. For a given $\mathbf{f} \in H^s(\Omega)^3$, let $\widehat{\mathbf{f}} = \iota^1(\mathbf{f}) = (D\mathbf{F})^T(\mathbf{f} \circ \mathbf{F})$. Since $\mathbf{F}$ is regular inside each element, we just need to check that the inter-element continuity is the one we expect. It is easy to see that, e.g.,

$$\frac{\partial \mathbf{F}}{\partial \zeta_1} \in \mathcal{H}^{s'}_{\mathbf{k}_1-1,\mathbf{k}_2,\mathbf{k}_3},\ \forall s' \in \mathbb{N},$$

and a similar result for the other partial derivatives implies that $\iota^1(\mathbf{f}) \in \mathcal{H}^{1,\mathbf{s}}$.

The inequalities follows by applying the chain rule. □

We are know ready to write the approximation estimate for the projectors Π^i, $i = 1, 2, 3$.

Theorem 6.3. *Let Assumption 2.7 and 5.1 hold. There exists a constant C depending only on $\mathbf{p}, \theta, \mathbf{F}$ such that for all elements $K = \mathbf{F}(Q) \in \mathcal{M}$, with $\widetilde{K} = \mathbf{F}(\widetilde{Q})$, it holds for $i = 0$ and $i = 3$:*

$$|f - \Pi^i f|_{H^r(K)} \le C h^{s-r}_{\widetilde{K}} \|f\|_{H^s(\widetilde{K})} \tag{6.4}$$

for all f in $H^s(\Omega)$, and r, s such that: when $i = 0$, $0 \le r \le s \le p + 1$; while when $i = 3$, $0 \le r \le s \le p_\ell$.

For $i = 1$ and $i = 2$ it holds:

$$|\mathbf{f} - \Pi^i \mathbf{f}|_{H^r(K)^3} \le C h^{s-r}_{\widetilde{K}} \|\mathbf{f}\|_{H^s(\widetilde{K})^3} \tag{6.5}$$

for all $\mathbf{f}$ in $H^s(\Omega)^3$, and r, s such that $0 \le r \le s \le p$.

Proof. We detail the proof of (6.5) for $i = 1$, but the reasoning is the same for all other estimates. Let $K \in \mathcal{M}$ and $Q = \mathbf{F}^{-1}(K)$. Using Proposition 6.2 and the definition of the projectors (6.1), for $\widehat{\mathbf{f}} = \iota^1(\mathbf{f})$ we have:

$$||\mathbf{f} - \Pi^1 \mathbf{f}||_{H^r(K)^3} \le C\|\widehat{\mathbf{f}} - \widehat{\Pi}^1 \widehat{\mathbf{f}}\|_{H^r(Q)^3}$$

Then, applying Proposition 4.9 and Proposition 6.2 again, and from the definition of the norms, we have, for $\mathbf{s}$: $|\mathbf{s}| \le s$, $0 \le r \le s \le p$,

$$\|\widehat{\mathbf{f}} - \widehat{\Pi}^1 \widehat{\mathbf{f}}\|_{H^r(Q)^3} \le C^{s-r}_{\widetilde{Q}} \|\mathbf{f}\|_{\mathcal{H}^{1,\mathbf{s}}(\widetilde{Q})}$$

Using the definition of the norms, the Proposition 6.2 and that $h_{\widetilde{Q}} \simeq h_{\widetilde{K}}$ thanks to Assumption 5.1, the result is proved. □

It is easy to see that:

Corollary 6.4. *Let Assumption 2.7 and 5.1 hold, and let $p = \min(p_1, p_2, p_3)$. Then there exists a constant C, only dependent on $\mathbf{p}, \theta, \mathbf{F}$ such that*

$$\begin{aligned}
\|f - \Pi^0 f\|_{H^r(\Omega)} &\leq C h^{s-r} \|f\|_{H^s(\Omega)} \qquad 0 \leq r \leq s \leq p+1, \\
\|\mathbf{f} - \Pi^1 \mathbf{f}\|_{H^r(\Omega)^3} &\leq C h^{s-r} \|\mathbf{f}\|_{H^s(\Omega)^3} \qquad 0 \leq r \leq s \leq p, \\
\|\mathbf{f} - \Pi^2 \mathbf{f}\|_{H^r(\Omega)^3} &\leq C h^{s-r} \|\mathbf{f}\|_{H^s(\Omega)^3} \qquad 0 \leq r \leq s \leq p, \\
\|f - \Pi^3 f\|_{H^r(\Omega)} &\leq C h^{s-r} \|f\|_{H^s(\Omega)} \qquad 0 \leq r \leq s \leq p,
\end{aligned}$$

for all $f \in H^s(\Omega)$ and $\mathbf{f} \in H^s(\Omega)^3$.

The approximation estimates in the previous corollary are presented in high order Sobolev norms. This gives in particular the error in the L^2 norm, which together with the commuting diagram property Lemma 4.8 will imply the approximation estimates in the graph norms for the spaces $\widehat{X}^i$.

Remark 6.5. *Following the previous remarks 4.5 and 4.10, it should be said that the same estimates hold when we replace Π_{p_ℓ} with $\widetilde{\Pi}_{p_\ell}$ and $\Pi^c_{p_\ell - 1}$ with $\widetilde{\Pi}^c_{p_\ell - 1}$, with some restrictions on exponents that is made clear in the next Proposition 6.7.*

We are now ready to turn to the "multipatch case", i.e., when $M_p > 1$. A complete approximation theory is beyond the scope of this paper, but the following observations maybe useful.

In particular, in order to define a projection operator locally patch by patch, we need to make sure that it matches at the patch interfaces, in a suitable sense.

Given a patch $\Omega^{(j)}$, let $\widetilde{\Pi}^{0,(j)}, \dots, \widetilde{\Pi}^{3,(j)}$ be the operators defined by the pull back relation (6.1), where $\widehat{\Pi}^0, \dots, \widehat{\Pi}^3$ are constructed starting from (4.7 - 4.10) but where we have replaced Π_{p_ℓ} with $\widetilde{\Pi}_{p_\ell}$ and $\Pi^c_{p_\ell - 1}$ with $\widetilde{\Pi}^c_{p-1}$, as defined in (2.16) and (2.25).

Let then

$$(\widetilde{\Pi}^i u)|_{\Omega^{(j)}} = \widetilde{\Pi}^{i,(j)}(u|_{\Omega^{(j)}}) \; i = 0, 3 \quad (\widetilde{\Pi}^i \mathbf{u})|_{\Omega^{(j)}} = \widetilde{\Pi}^{i,(j)}(\mathbf{u}|_{\Omega^{(j)}}) \; i = 1, 2.$$

The following holds:

Proposition 6.6. *Let Assumptions 5.1 and 5.2 hold. The operators $\widetilde{\Pi}^i$, $i = 0, \dots, 3$ map regular functions onto $X^i_h(\Omega)$, $i = 0, \dots, 3$ and verify the following property:*

$$\begin{array}{ccccccccccc}
\mathbb{R} & \longrightarrow & X^0_{\mathrm{reg}} & \xrightarrow{\mathbf{grad}} & X^1_{\mathrm{reg}} & \xrightarrow{\mathbf{curl}} & X^2_{\mathrm{reg}} & \xrightarrow{\mathrm{div}} & X^3_{\mathrm{reg}} & \longrightarrow & 0 \\
 & & \widetilde{\Pi}^0 \downarrow & & \widetilde{\Pi}^1 \downarrow & & \widetilde{\Pi}^2 \downarrow & & \widetilde{\Pi}^3 \downarrow & & \\
\mathbb{R} & \longrightarrow & X^0_h(\Omega) & \xrightarrow{\mathbf{grad}} & X^1_h(\Omega) & \xrightarrow{\mathbf{curl}} & X^2_h(\Omega) & \xrightarrow{\mathrm{div}} & X^3_h(\Omega) & \longrightarrow & 0
\end{array} \tag{6.6}$$

where X^i_{reg} stands for $X^i \cap C^\infty(\Omega)$, $i = 0, \dots 3$.

Proof. From the very definition of $\widetilde{\Pi}_{p,\Xi}$ in (2.16) it is clear that it is interpolatory at the endpoints of the interval. In view of the tensorization and the definition of (4.7 - 4.10), $(\widetilde{\Pi}^{0,(j)} u)|_{\Gamma_{ij}}$ depends only on $u|_{\Gamma_{ij}}$, when Γ_{ij} is a face of $\Omega^{(j)}$. Analogously, $(\widetilde{\Pi}^{1,(j)} \mathbf{u} \times \mathbf{n}_{ij})|_{\Gamma_{ij}}$ ($\mathbf{n}_{ij}$ being a normal to Γ_{ij}) depends only on $\mathbf{u}|_{\Gamma_{ij}} \times \mathbf{n}_{ij}$ and $(\widetilde{\Pi}^{2,(j)} \mathbf{u} \cdot \mathbf{n}_{ij})|_{\Gamma_{ij}}$ depends only upon $\mathbf{u}|_{\Gamma_{ij}} \cdot \mathbf{n}_{ij}$. □

Moreover, following the Remarks 4.5 and 4.10, by definition of $\widetilde{\Pi}^i$, the following error estimates holds:

Proposition 6.7. *Let Assumptions 2.7, 5.1 and 5.2 hold, and let* $p = \min(p_1, p_2, p_3)$. *Then there exists a constant* C, *only dependent on* $\mathbf{p}, \theta, \mathbf{F}$ *such that*

$$\begin{aligned}
\|f - \widetilde{\Pi}^0 f\|_{H^r(\Omega)} &\leq Ch^{s-r}\|f\|_{H^s(\Omega)} && 0 \leq r \leq s \leq p+1, s > 3/2+\epsilon \\
\|\mathbf{f} - \widetilde{\Pi}^1 \mathbf{f}\|_{H^r(\Omega)^3} &\leq Ch^{s-r}\|\mathbf{f}\|_{H^s(\Omega)^3} && 0 \leq r \leq s \leq p, s > 1+\epsilon \\
\|\mathbf{f} - \widetilde{\Pi}^2 \mathbf{f}\|_{H^r(\Omega)^3} &\leq Ch^{s-r}\|\mathbf{f}\|_{H^s(\Omega)^3} && 0 \leq r \leq s \leq p, s > \epsilon \\
\|f - \widetilde{\Pi}^3 f\|_{H^r(\Omega)} &\leq Ch^{s-r}\|f\|_{H^s(\Omega)} && 0 \leq r \leq s \leq p,
\end{aligned}$$

for all $f \in H^s(\Omega)$ *and* $\mathbf{f} \in H^s(\Omega)^3$, *where* ϵ *stands for an arbitrarily small, but positive, number.*

It should be said that this result is not satisfactory because L^2 stability would be desirable. The study of L^2 stable commuting projectors is left to future works.

7. Conclusions

We have presented here the construction and the main properties of the spline complex. Many other mathematical properties may be studied, as, e.g., following Remark 4.4, the construction of chain / co-chain dualities that may be useful for the discretization of some partial differential equations. At today, these spaces has been applied in various contexts, starting from the most classical one, i.e., the Maxwell equations [2, 3, 16, 18], and, more generally, could be applied to the discretization of the Hodge Laplacian (see e.g., [9]). Thanks to the regularity of the spline representation for vector fields, the spline complex can be exploited also in contexts where the finite element complex would not fit. E.g., it was used to treat Stokes and Navier-Stokes equations [19–22], and to simulate Reissner-Mindlin plates [23]. Other applications are possible and are currently under study.

Finally, the spline complex provides representation of tangential vector fields on manifolds that have the same regularity as the manifold itself, and this fact maybe useful also in contexts that are far from the discretization of partial differential equations.

References

[1] A. Buffa, G. Sangalli, and R. Vázquez, *Isogeometric methods for computational electromagnetics: B-spline and T-spline discretizations*, J. Comput. Phys. **257**, Part B (2014), 1291–1320.

[2] A. Buffa, J. Rivas, G. Sangalli, and R. Vázquez, *Isogeometric discrete differential forms in three dimensions*, SIAM J. Numer. Anal. **49**(2) (2011), 818–844.

[3] A. Buffa, G. Sangalli, and R. Vázquez, *Isogeometric analysis in electromagnetics: B-splines approximation*, Comput. Methods Appl. Mech. Engrg. **199**(17-20) (2010), 1143–1152.

[4] L. Beirão da Veiga, A. Buffa, G. Sangalli, and R. Vázquez, *Mathematical analysis of variational isogeometric methods*, Acta Numerica **23** (2014), 157–287.

[5] T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs, *Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement*, Comput. Methods Appl. Mech. Engrg. **194**(39-41) (2005), 4135–4195.

[6] J. A. Cottrell, T. J. R. Hughes, and Y. Bazilevs, *Isogeometric Analysis: toward integration of CAD and FEA*, John Wiley & Sons, 2009.

[7] L. Piegl and W. Tiller, *The Nurbs Book*, Springer-Verlag, New York, 1997.

[8] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer. **11** (2002), 237–339.

[9] D. N. Arnold, R. S. Falk, and R. Winther, *Finite element exterior calculus, homological techniques, and applications*, Acta Numer. **15** (2006), 1–155.

[10] D. Boffi, *Finite element approximation of eigenvalue problems*, Acta Numer. **19** (2010), 1–120.

[11] D. N. Arnold, *Differential complexes and numerical stability*, in: Proceedings of the International Congress of Mathematicians, Vol. 1, 2002, pp. 137–157.

[12] L. L. Schumaker, *Spline functions: basic theory*, 3rd Edition, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 2007.

[13] C. de Boor, *A practical guide to splines*, revised Edition, Vol. 27 of Applied Mathematical Sciences, Springer-Verlag, New York, 2001.

[14] Y. Bazilevs, L. Beirão da Veiga, J. A. Cottrell, T. J. R. Hughes, and G. Sangalli, *Isogeometric analysis: approximation, stability and error estimates for h-refined meshes*, Math. Models Methods Appl. Sci. **16**(7) (2006), 1031–1090.

[15] P. Monk, *Finite element methods for Maxwell's equations*, Oxford University Press, Oxford, 2003.

[16] A. Ratnani and E. Sonnendrücker, *An arbitrary high-order spline finite element solver for the time domain Maxwell equations*, J. Sci. Comput. **51** (2012), 87–106.

[17] A. Buffa and S. H. Christiansen, *A dual finite element complex on the barycentric refinement*, Math. Comp. **76**(260) (2007), 1743–1769 (electronic).

[18] R. Vázquez, A. Buffa, and L. Di Rienzo, *Isogeometric FEM implementation of high order surface impedance boundary conditions*, IEEE Trans. Magn.To appear. doi:10.1109/TMAG.2014.2298435.

[19] A. Buffa, C. de Falco, and G. Sangalli, *Isogeometric Analysis: stable elements for the 2D Stokes equation*, Internat. J. Numer. Methods Fluids **65**(11-12) (2011), 1407–1422.

[20] J. A. Evans and T. J. R. Hughes, *Isogeometric divergence-conforming B-splines for the Darcy-Stokes-Brinkman equations.*, Math. Models Methods Appl. Sci. **23**(04) (2013), 671–741.

[21] _____, *Isogeometric divergence-conforming B-splines for the Steady Navier-Stokes Equations*, Math. Models Methods Appl. Sci. **23**(08) (2013), 1421–1478.

[22] _____, *Isogeometric divergence-conforming B-splines for the Unsteady Navier-Stokes Equations*, J. Comput. Phys. **241** (2013), 141 – 167.

[23] L. Beirão da Veiga, A. Buffa, C. Lovadina, M. Martinelli, and G. Sangalli, *An isogeometric method for the Reissner-Mindlin plate bending problem*, Comput. Methods Appl. Mech. Engrg. **209**-212 (2012), 45 – 53.

Istituto di Matematica Applicata e Tecnologie Informatiche "E. Magenes", Via Ferrata 1, 27100 Pavia, Italy

E-mail: annalisa.buffa@imati.cnr.it

Multiscale model reduction with generalized multiscale finite element methods

Yalchin Efendiev

Abstract. Many application problems have multiscale nature. Due to disparity of scales, the simulations of these problems are prohibitively expensive. Some types of upscaling or model reduction techniques are needed to solve many multiscale problems. In this paper, we discuss a few known techniques that are used for problems with scale separation and focus on Generalized Multiscale Finite Element Method (GMsFEM) that has been recently proposed for solving problems with non-separable scales and high contrast. The main objective of the method is to provide local reduced-order approximations for linear and nonlinear PDEs via multiscale spaces on a coarse computational grid. In the paper, we briefly discuss some main concepts of constructing multiscale spaces and applications of GMsFEMs.

Mathematics Subject Classification (2010). Primary 65N99; Secondary 65N30.

Keywords. Multiscale, finite element, porous media, homogenization, model reduction.

1. Introduction

Many problems involve media or processes that contain multiple scales and physical properties that vary over orders of magnitude and exhibit uncertainties. As an example, solutions to fluid flow problems in heterogeneous porous media require large-scale computations to understand the complex physics and chemistry occurring in the subsurface. These models, henceforth called fine-grid models, often consist of over $10^6 - 10^7$ gridcells. The ability to *coarsen* these highly resolved models to levels of detail appropriate for simulations, optimization, and uncertainty quantification, while maintaining the integrity of the model for its fast simulation is clearly needed. Similarly, complexity makes many other, e.g., seismic, hydrological applications to be computationally challenging. Traditional methods share a common potential weakness in that the computational cost will be prohibitively large for many of these multiscale problems.

There are a variety of model reduction techniques that include homogenization, upscaling, perturbation approaches, multiscale methods, global model reduction techniques, and so on. Some of them that are closely related to the proposed methods includes homogenization, numerical upscaling, and multiscale finite element methods. Upscaling techniques (see e.g., [12, 13]) have been commonly used in many applications and include the re-formulation of the global problem on a coarse grid which are called upscaled equations. The upscaled equations contain effective media properties. The calculations of these effective properties typically involve solving local problems in representative volumes or in coarse-grid blocks and

Proceedings of the International Congress of Mathematicians, Seoul, 2014

extracting these properties via volume averaging. Though effective in many cases, these approaches do not systematically approximate the fine-grid solution. Some recent approaches introduced in the context of Multiscale Finite Element Methods (see [1–5, 8, 25–27, 29, 37–39, 43]) that can systematically and effectively enrich the solution space locally on a coarse-grid level. The main idea of these multiscale methods is to construct an approximation space for the solution on each coarse (computational) grid. Many of these approaches have focused on finding a limited number of basis functions for approximating the solution space. In this paper, we discuss the recently introduced Generalized Multiscale Finite Element Method ([21]) that attempts to systematically identify local basis functions

In the paper, we give a brief overview of multiscale model reduction methods. We start with the problems that contain scale separation and discuss homogenization and numerical homogenization procedures. Furthermore, we discuss multiscale finite element methods that use one basis function per coarse-element node, and then introduce Generalized Multiscale Finite Element Method.

The Generalized Multiscale Finite Element Method (GMsFEM) is a flexible framework that generalizes the Multiscale Finite Element Method (MsFEM) by systematically enriching the coarse spaces and taking into account small scale information and complex input spaces. This approach, as in many multiscale model reduction techniques, divides the computation into two stages: offline and online. In the offline stage, a small dimensional space is constructed that can be efficiently used in the online stage to construct multiscale basis functions. These multiscale basis functions can be re-used for any input parameter to solve the problem on a coarse grid. The main idea behind the construction of offline and online spaces is the selection of local spectral problems and the selection of the snapshot space. We briefly discuss how the method can be used within different global finite element discretizations and applied for various applications. This paper is not intended to give many details of the method, its implementation and its applications which can be found in the literature.

2. Preliminaries

2.1. A model problem. Throughout the paper, we will consider a model problem that describes flow in heterogeneous media. The governing equations are given by (subject to some boundary conditions)

$$-\operatorname{div}\big(\kappa(x)\,\nabla u\big) = -\frac{\partial}{\partial x_i}\left(\kappa_{ij}(x)\frac{\partial}{\partial x_j}u\right) = f, \tag{2.1}$$

where $\kappa(x)$ is assumed to be multiscale field representing the media properties. The summation over repeated indices is assumed throughout. The methods discussed in the paper are applicable to a wide range of problems. One of main emphasis of the paper is on handling multiscale features of the solution space locally.

2.2. Scales. We briefly describe some multiscale heterogeneities. One of simpler multiscale functions that are often used in designing multiscale methods have scale separation. An example is a two-scale function

$$\kappa(x) = \kappa(x, \frac{x}{\epsilon}),$$

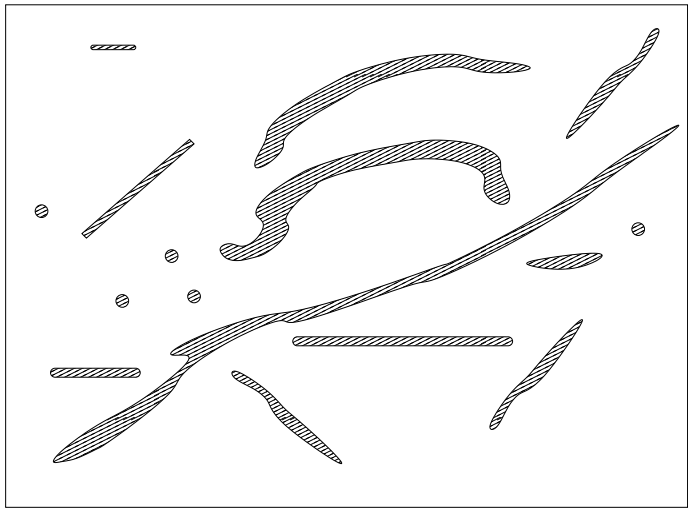


Figure 2.1. Schematic illustration of high-contrast heterogeneous features

where $\epsilon \ll 1$. Fast variable is typically denoted by

$$y = \frac{x}{\epsilon}.$$

In this case, we can write $\kappa = \kappa(x, y)$, where $x \in R^d$ and $y \in R^d$. Such functions vary on distinct scales.

The multiscale nature of the function can be made more complicated by introducing additional scales. For example, one can use functions of different frequencies to form a function of multiple frequencies

$$\kappa_\epsilon(x) = \sum_i a_i g_i(x, \frac{x}{r_i \epsilon}),$$

where $g_i(x, y)$ are periodic with respect to y and r_i are incommensurable numbers that make the functions be non-periodic. Such functions can still have scale separation while span a variety of scales within a range.

In this paper, we will deal with spatial fields (that appear as coefficients in PDEs) that do not have scale separation and contain high contrast. Such permeability fields arise in many porous media applications in higher dimensions. For example, by introducing complex fine-scale features in 2D that have irregular shapes with small width and various shapes. Such features can represent high or low conductivity fields in the subsurface and represent an example when one can have many scales. Moreover, the media properties (such as conductivity field) within these features can be much larger (or lower) than the background permeability field (such that the ratio in the contrast is of order small length scales), see Figure 2.1 for a schematic illustration. In this case, one deals with high-contrast permeability fields and has to exercise a caution when taking the limit with respect to the spatial length scales.

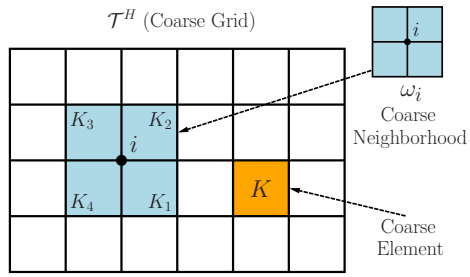


Figure 2.2. Schematic of a coarse element and coarse neighborhood

2.3. Coarse and fine mesh description. To describe the general solution framework for the model equations in this paper, we first introduce the notion of fine and coarse grids. We assume that the problem under consideration can be solved on a fine grid denoted by $\mathcal{T}^h$, where h is the fine-mesh size. The fine grid typically consists of usual conforming partition of the computational domain D into finite elements (triangles, quadrilaterals, tetrahedrals, etc.). Our objective is to avoid performing the computations on the fine grid and reduce the problem to solving it on a coarse grid.

We let $\mathcal{T}^H$ be a coarse grid and assume that each coarse subregion is partitioned into a connected union of fine grid blocks. Here, H denotes the coarse-mesh size. We use $\{x_i\}_{i=1}^{N_v}$ (where N_v the number of coarse nodes) to denote the vertices of the coarse mesh $\mathcal{T}^H$, and define the neighborhood of the node x_i by

$$\omega_i = \bigcup \{K_j \in \mathcal{T}^H; \quad x_i \in \overline{K}_j\}. \tag{2.2}$$

See Fig. 2.2 for an illustration of neighborhoods and elements subordinated to the coarse discretization. We emphasize the use of ω_i to denote a coarse neighborhood, and K to denote a coarse element.

3. Some existing local model reduction techniques

In this section, we present some model reduction techniques (that are relevant to the methods discussed in the paper) that use a fixed number of degrees of freedom in each coarse patch. Traditional homogenization and numerical homogenization techniques employ a few degrees of freedom to approximate the solution space in each patch. In a similar fashion, one can consider multiscale finite element methods that use only one (or a fixed number of) basis function per coarse node. In our next section, we will discuss Generalized Multiscale Finite Element Methods (GMsFEM) that is a general strategy of identifying local approximation space in each coarse patch for multiscale problems.

We consider a model problem (2.1) and later discuss generalizations to selected applications.

3.1. Homogenization. In this section, we present a low-order homogenization expansion for the solution of elliptic PDE and emphasize low cost computational approximation based on this expansion.

We assume (see (2.1))

$$\kappa(x) = \kappa(x, \frac{x}{\epsilon}).$$

One can show (see [40]) that the solution of (2.1) can be approximated by

$$\widehat{u}_\epsilon(x) = u_0(x) + \epsilon\chi_i(x, \frac{x}{\epsilon})\frac{\partial}{\partial x_i}u_0, \tag{3.1}$$

where $\chi_l(x, y)$ (periodic in y) solves

$$\frac{\partial}{\partial y_i}\left(\kappa_{ij}(x, y)\frac{\partial}{\partial y_j}\chi_l(x, y)\right) = -\frac{\partial}{\partial y_i}\kappa_{il}(x, y). \tag{3.2}$$

Here u_0 is the homogenized solution that satisfies (subject to some boundary conditions)

$$-\frac{\partial}{\partial x_i}\left(\kappa^*_{ij}(x)\frac{\partial}{\partial x_j}u_0(x)\right) = f(x),$$

with the homogenized coefficients

$$\kappa^*_{ij}(x) = \frac{1}{|Y|}\int_Y \kappa_{il}(x, y)\left(\delta_{jl} + \frac{\partial}{\partial y_l}\chi_j(x, y)\right)dy,$$

where Y is the unit period. It can be shown that [40]

$$\|u_\epsilon - \widehat{u}_\epsilon\|_{H^1(D)} \leq C\sqrt{\epsilon}, \tag{3.3}$$

where D is the domain.

The above expansion (3.1) shows that the solution of the multiscale PDE (2.1) can be approximated in each coarse patch using the solution of cell problems (3.2). We only need d degrees of freedom (represented by functions χ_i, $i = 1, ..., d$) to approximate the local fine-scale features of the solution in R^d. The error of this approximation can be estimated based on (3.3). We note that many numerical methods have been developed for solving problems with scale separation (e.g., [6, 19, 41, 42]) that we do not discuss here. The cost of solving the homogenized problem and the cell problems are independent of ϵ.

3.2. Numerical homogenization. Numerical homogenization is often based on homogenization and approximates the effective (homogenized) coefficients and the solutions of the homogenized equations in a numerical way. Though these approaches are based on homogenization, they are often used for cases without periodicity and even with no scale separation (e.g., in subsurface applications, see [18]). Below, we briefly present a numerical homogenization for our model problem.

The main idea of numerical homogenization is to identify the homogenized coefficients in each coarse-grid block. The basic underlying principle is to compute the upscaled quantities such that they preserve some averages for a given set of local boundary conditions.

We again consider (2.1) though these methods can be applied to various linear and nonlinear problems. Our objective is to define an upscaled (or homogenized) conductivity for each coarse block, in general without assuming periodicity. We follow homogenization technique and solve local problems for each coarse block subject to some boundary condition

$$\frac{\partial}{\partial x_i}\left(\kappa_{ij}(x)\frac{\partial}{\partial x_j}\phi_l\right) = 0 \text{ in } K, \tag{3.4}$$

where K is a coarse block (see Figure 2.2 for illustration). The choice of boundary conditions is important and various boundary conditions can be used (see e.g., [44]). For example, Dirichlet boundary conditions are often used. In this case,

$$\phi_l = x_l \text{ on } \partial K.$$

The other choice for boundary condition is periodic boundary condition. In this case,

$$\phi_l = x_l + \text{periodic function on } \partial K.$$

The upscaled coefficients, $\kappa_{ij}^{*,nh}$, are defined by averaging the fluxes:

$$\int_K \kappa_{ij}^{*,nh} \frac{\partial}{\partial x_j}\phi_l^* = \int_K \kappa_{ij}(x)\frac{\partial}{\partial x_j}\phi_l. \tag{3.5}$$

The motivation behind this upscaling is to state that the average flux response for the fine-scale local problem with prescribed boundary conditions is the same as that for the upscaled solution. Now, if we take $\phi_l^* = x_l$ in K, we have

$$\kappa_{il}^{*,nh} = \frac{1}{|K|}\int_K \kappa_{ij}(x)\frac{\partial}{\partial x_j}\phi_l. \tag{3.6}$$

Once the homogenized coefficients are computed, the upscaled solution is found by solving

$$\frac{\partial}{\partial x_i}\left(\kappa_{ij}^{*,nh}(x)\frac{\partial}{\partial x_j}u^*(x)\right) = f. \tag{3.7}$$

The proximity of u and u^* can be shown in the case of scale separation [44]. The error can be reduced by using larger domains [44] and computing the effective permeabilities.

3.3. Multiscale Finite Element Method (MsFEM). In this section, we briefly present MsFEM that uses one basis function per coarse-element node. This method (as presented in this section) shares similarities with numerical upscaling techniques as it uses only one basis function per coarse node. However, there are a number of important advantages of using MsFEMs (see [27] for more details). We will briefly mention them at the end of the section.

MsFEM basically consists of two parts

- basis function construction
- a choice of the global formulation that couples these basis functions.

First, we discuss the basis function construction. Let ϕ_i^0 be the nodal basis of the standard finite element space W_H on a coarse grid $\mathcal{T}^H$, $W_H \subset H_0^1(\Omega)$. Denote by S_i the support of ϕ_i^0 and define ϕ_i with support in S_i as follows

$$L\phi_i = 0 \text{ in } K, \quad \phi_i = \phi_i^0 \text{ on } \partial K, \ \forall\, K \in \mathcal{T}^H,\ K \subset S_i, \tag{3.8}$$

where L is the linear elliptic operator that corresponds to (2.1). Note that even though the choice of ϕ_i^0 can be quite arbitrary, our main assumption is that the basis functions satisfy the leading order homogeneous equations when the right hand side f is a smooth function. We would like to remark that MsFEM formulation allows one to take an advantage of scale separation. In particular, K can be chosen to be a volume smaller than the coarse

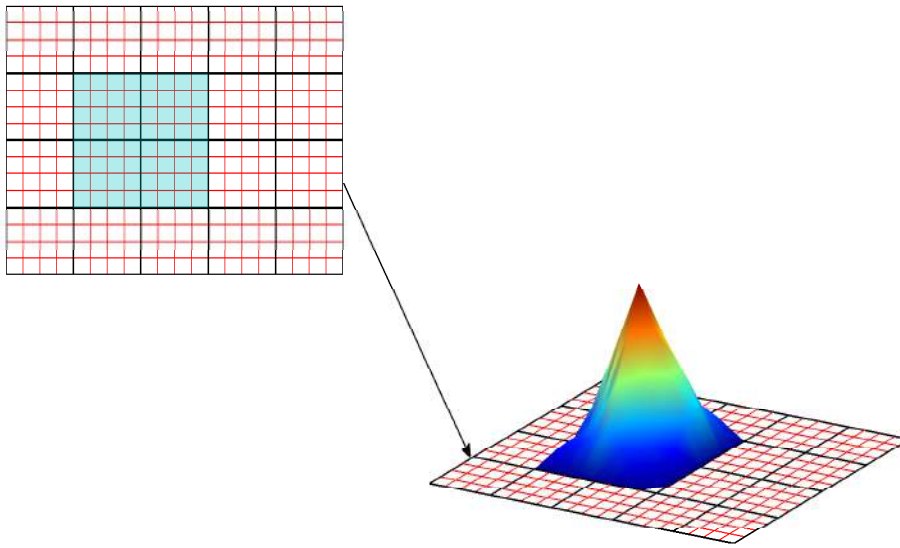


Figure 3.1. An illustration of two dimensional multiscale basis functions

grid. Indeed, in the presence of scale separation, one can use the solution in Representative Volume Element (RVE) to represent the solution in the entire region as it is done in classical homogenization. Once the basis functions are constructed, we let V_H be the finite element space spanned by ϕ_i.

In the above discussion, we presented simplest basis function construction and a global formulation. In general, the global formulation can be easily modified and various global formulations based on finite volume, mixed finite element, discontinuous Galerkin finite element and other methods can be derived. One can also consider the applications of these techniques to nonlinear problems [27].

3.3.1. Boundary conditions for basis functions. As for basis functions, the choice of boundary conditions in defining the multiscale basis functions plays a crucial role in approximating the multiscale solution. Intuitively, the boundary condition for the multiscale basis function should reflect the multiscale oscillation of the solution u across the boundary of the coarse grid element. By choosing a linear boundary condition for the basis function, we will create a mismatch between the exact solution u and the finite element approximation across the element boundary. This issue is studied in the literature (e.g., [28]) and an oversampling technique (see e.g., [28, 43]) is introduced to alleviate this difficulty. This technique enables us to remove the artificial numerical boundary layer across the coarse grid boundary element.

We note that MsFEM has several advantages over numerical homogenization techniques. Some of them include: (1) fine-scale recovery of the solution based on multiscale basis functions; (2) imposing important global information on multiscale basis functions; (3) flexible coarse gridding; (4) the use of enrichment which will be discussed next.

4. Generalized Multiscale Finite Element Methods

In this section, we discuss main ingredients of GMsFEM such as snapshot space construction, local basis construction, some selected global coupling mechanisms for multiscale basis functions, and the applications.

4.1. Basic concept. In this section we will describe some details associated with an offline-online procedure for constructing GMsFEM basis functions on a model problem (2.1). We note that this procedure is applicable for the general case(s) when the coefficient of a system depends on a parameter μ. That is, we may assume that $\kappa = \kappa(x;\mu)$ for the model problems that we consider. A general outline for the procedure is offered below.

1. Offline computations:
   - 1.0. Coarse grid generation.
   - 1.1. Construction of snapshot space that will be used to compute an offline space.
   - 1.2. Construction of a small dimensional offline space by performing dimension reduction in the space of local snapshots.
2. Online computations:
   - 2.1. For each input parameter, compute multiscale basis functions.
   - 2.2. Solution of a coarse-grid problem for any force term and boundary condition.
   - 2.3. Iterative solvers, if needed.

4.2. Local basis functions. In the offline computation, we first construct a snapshot space V^{τ}_{snap}, corresponding to either the continuous Galerkin (CG) or discontinuous Galerkin (DG) formulation. Construction of the snapshot space involves solving the local problems for various choices of input parameters, on a specified coarse subdomain τ, where τ denotes coarse neighborhood-based computations for a CG formulation (ω_i), and coarse element-based computations (K) for a DG formulation [23]. For brevity of notation we now omit the superscript τ when dealing with local problems, yet it is assumed throughout this section that the offline and online space computations are localized to respective coarse subdomains.

The choice of the snapshot space depends the global discretization and the particular application. The choice snapshot space helps (1) achieving faster convergence rate (2) imposing problem relevant restriction on the coarse spaces (such as divergence free elements, and so on) (3) reducing the computational cost of calculating offline spaces. We refer to a number of papers in the literature where various snapshot spaces are considered. For our model problem (Eq. (2.1)), one can consider the snapshot space to be (1) harmonic extensions (2) local fine-grid functions, (3) dominant eigenvectors of local eigenvalue problems or other choices. The snapshot space generated by harmonic extensions is constructed by solving local problems

$$L_{\mu_j}(\psi^{\tau}_{l,j}) = 0 \text{ in } \tau \tag{4.1}$$

with boundary conditions

$$\psi^{\tau}_{l,j} = b_l \text{ in } \partial\tau, \tag{4.2}$$

with b_l being selected shape or basis functions along the boundary $\partial\tau$. Here, L_μ is our model problem (2.1) with $\kappa = \kappa(x,\mu)$ and $\{\mu_j,\ \text{for } j = 1,\cdots,J\}$ is the selected parameters for generation of snapshots and there are various ways to generate this set (see e.g., [21]) One can also consider local fine-grid functions or dominant eigenvectors for local eigenvalue problems possibly formulated in oversampled regions [23]. For the numerical implementation, we reorder the snapshot functions using a single index to create the matrix

$$R_{\text{snap}} = \left[\psi_1^{\text{snap}}, \ldots, \psi_{M_{\text{snap}}}^{\text{snap}}\right],$$

where M_{snap} denotes the total number of functions to keep in the snapshot matrix construction.

In order to construct the offline space V_{off}^τ, we perform a dimension reduction of the snapshot space using an auxiliary spectral decomposition. The main objective is to use the offline space to construct a set of multiscale basis functions for each μ value in the online stage. At the offline stage the bilinear forms are chosen to be *parameter-independent*, such that there is no need to reconstruct the offline space for each μ value. The choice of this local eigenvalue problem is motivated by the analysis and depends on several factors that include (1) the global formulation (2) the smoothness of the solution (3) the behavior of the eigenvalues and so on.

We seek the subspace V_{off}^τ such that for any μ and $\psi \in V_{\text{snapshots}}^\tau(\mu)$ ($V_{\text{snapshots}}^\tau(\mu)$ is the space of snapshots which are computed for a given μ), there exists $\psi_0 \in V_{\text{off}}^\tau$, such that, for all μ,

$$a_\tau^{\text{off}}(\psi - \psi_0, \psi - \psi_0; \mu) \preceq \delta s_\tau^{\text{off}}(\psi - \psi_0, \psi - \psi_0; \mu), \tag{4.3}$$

where $a_\tau^{\text{off}}(\phi,\phi;\mu)$ and $s_\tau^{\text{off}}(\phi,\phi;\mu)$ are auxiliary bilinear forms that are motivated by the analysis. In computations, this involves solving an eigenvalue problem with a mass matrix and the basis functions are selected based on dominant eigenvalues. As we pointed out earlier the choice of this eigenvalue problem is motivated by the analysis and depends on the discretization that is used to couple basis functions and the underlying problem. In the discrete setup, this involves solving local eigenvalue problems in the snapshot space. In the case of our model problem, one can use

$$A^{\text{off}}\Psi_k^{\text{off}} = \lambda_k^{\text{off}} S^{\text{off}}\Psi_k^{\text{off}}, \tag{4.4}$$

where

$$A^{\text{off}} = [a_{mn}^{\text{off}}] = \int_\tau \kappa(x,\overline{\mu})\nabla\psi_m^{\text{snap}} \cdot \nabla\psi_n^{\text{snap}} = R_{\text{snap}}^T \overline{A} R_{\text{snap}}$$

$$S^{\text{off}} = [s_{mn}^{\text{off}}] = \int_\tau \widetilde{\kappa}(x,\overline{\mu})\psi_m^{\text{snap}}\psi_n^{\text{snap}} = R_{\text{snap}}^T \overline{S} R_{\text{snap}},$$

where $\kappa(x,\overline{\mu})$, and $\widetilde{\kappa}(x,\overline{\mu})$ are domain-based averaged coefficients with $\overline{\mu}$ chosen as the average of pre-selected μ_i's, and the form for $\widetilde{\kappa}$ can be found in [21].

For a given input parameter, we next construct the associated online coarse space $V_{\text{on}}^\tau(\mu)$ *for each* μ value on each coarse subdomain. Note that for parameter-independent case, there is no need for the online space and one uses the offline space to solve the problem. The online coarse space will be used within the finite element framework to solve the original global problem, where a continuous or discontinuous Galerkin coupling of the multiscale basis functions is used to compute the global solution. In particular, we seek a subspace of the offline space such that it can approximate any element of the offline space in an appropriate

sense. We note that at the online stage, the bilinear forms are chosen to be *parameter-dependent*. For each τ and for each input parameter, we will formulate a quotient for finding a subspace of $V^\tau_{\text{on}}(\mu)$ where the space will be constructed for each μ (independent of source terms). We seek a subspace $V^\tau_{\text{on}}(\mu)$ of V^τ_{off} such that for each $\psi \in V^\tau_{\text{off}}$, there exists $\psi_0 \in V^\tau_{\text{on}}(\mu)$ such that

$$a^{\text{on}}_\tau(\psi-\psi_0,\psi-\psi_0;\mu) \preceq \delta s^{\text{on}}_\tau(\psi-\psi_0,\psi-\psi_0;\mu) \tag{4.5}$$

for some prescribed error tolerance δ (different from the one in the offline stage), and the choices of $a^{\text{on}}_{\omega_i}$ and $s^{\text{on}}_{\omega_i}$ that comes from the analysis. As before, the choice of the local eigenvalue problem is motivated by the analysis and depends on the global discretization and on the offline spaces. In a discrete setup, the following eigenvalue problem for our model problem is solved

$$A^{\text{on}}(\mu)\Psi^{\text{on}}_k = \lambda^{\text{on}}_k S^{\text{on}}(\mu)\Psi^{\text{on}}_k, \tag{4.6}$$

where

$$A^{\text{on}}(\mu) = [a^{\text{on}}(\mu)_{mn}] = \int_\tau \kappa(x;\mu)\nabla\psi^{\text{off}}_m\cdot\nabla\psi^{\text{off}}_n = R^T_{\text{off}}A(\mu)R_{\text{off}}$$

$$S^{\text{on}}(\mu) = [s^{\text{on}}(\mu)_{mn}] = \int_\tau \widetilde{\kappa}(x;\mu)\psi^{\text{off}}_m\psi^{\text{off}}_n = R^T_{\text{off}}S(\mu)R_{\text{off}},$$

and $\kappa(x;\mu)$ and $\widetilde{\kappa}(x;\mu)$ are now parameter dependent. To generate the online space, we then choose the smallest M_{on} eigenvalues from Eq. (4.6) and form the corresponding eigenvectors in the offline space by setting $\psi^{\text{on}}_k = \sum_j \Psi^{\text{on}}_{kj}\psi^{\text{off}}_j$ (for $k = 1,\dots,M_{\text{on}}$), where Ψ^{on}_{kj} are the coordinates of the vector ψ^{on}_k. We note that in the case when the coefficient is independent of the parameter, then $V_{\text{on}} = V_{\text{off}}$. In other words, the online space discussion is limited to the case where the coefficient is parameter-dependent.

4.3. Global coupling.

4.3.1. Galerkin coupling. For a conforming Galerkin formulation, we need conforming basis functions, and τ denote ω_i, as defined in Eqn. (2.2) and shown in Fig. 2.2. We modify V^τ_{on} by multiplying the functions from this space with partition of unity functions. The modified space has the same dimension and is given by $\text{Span}_j(\chi_i\psi^{\tau,\text{on}}_j)$, where $\psi^{\tau,\text{on}}_j \in V^\tau_{\text{on}}(\mu)$ and χ_i is supported in τ. Then, the Galerkin approximation can be written as

$$u^G_{\text{ms}}(x;\mu) = \sum_{i,j} c^i_j\chi_i(x)\psi^{\tau,\text{on}}_j(x;\mu).$$

If we introduce

$$V^G_{\text{on}} = \text{Span}_{i,j}(\chi_i\psi^{\tau,\text{on}}_j), \tag{4.7}$$

then Galerkin formulation is given by

$$\kappa(u^G_{ms},v;\mu) = (f,v),\ \forall\, v\in V^G_{\text{on}}, \tag{4.8}$$

where $\kappa(u,v;\mu)$ corresponds to the bilinear form associated with (2.1) with $\kappa=\kappa(x,\mu)$ and (f,v) is the usual L^2-inner product.

4.3.2. Discontinuous Galerkin coupling. One can also use the discontinuous Galerkin (DG) approach, in particular, interior penalty DG, to couple multiscale basis functions. This may avoid the use of the partition of unity functions; and the local coarse region τ denotes the coarse block K as depicted in Fig. 2.2, however, a global formulation needs to be chosen carefully. We omit the parameter μ in the global formulation description. The global formulation is given by

$$\kappa^{DG}(u,v) = f(v) \quad \text{for all} \quad v = \{v_\tau \in V_{\text{on}}^\tau\}, \tag{4.9}$$

where

$$\kappa^{DG}(u,v) = \sum_\tau \kappa_\tau^{DG}(u,v) \quad \text{and} \quad f(v) = \sum_\tau \int_\tau f v_\tau dx \tag{4.10}$$

for all $u = \{u_\tau\}, v = \{v_\tau\}$. Each local bilinear form κ_τ^{DG} is given as a sum of three bilinear forms:

$$\kappa_\tau^{DG}(u,v) := \kappa_\tau(u,v) + r_\tau(u,v) + p_\tau(u,v), \tag{4.11}$$

where κ_τ is the bilinear form,

$$\kappa_\tau(u,v) := \int_\tau \kappa_r \nabla u_\tau \cdot \nabla v_\tau dx, \tag{4.12}$$

where κ_r is the restriction of $\kappa(x)$ in τ; the r_τ is the symmetric bilinear form,

$$r_\tau(u,v) := \sum_{E \subset \partial\tau} \frac{1}{l_E} \int_E \widetilde{\kappa}_E \left(\frac{\partial u_\tau}{\partial n_\tau}(v_\tau - v_{\tau'}) + \frac{\partial v_\tau}{\partial n_\tau}(u_{\tau'} - u_\tau) \right) ds,$$

where $\widetilde{\kappa}_E$ is a weighted average of $\kappa(x)$ near the edge E, l_E is the length of the edge E, and τ' and τ are two coarse-grid elements sharing the common edge E; and p_τ is the penalty bilinear form,

$$p_\tau(u,v) := \sum_{E \subset \partial K} \frac{1}{l_E} \delta_E \int_E \widetilde{\kappa}_E (u_{\tau'} - u_\tau)(v_{\tau'} - v_\tau) ds. \tag{4.13}$$

Here δ_E is a positive penalty parameter that needs to be selected and its choice affects the performance of GMsFEM. One can choose eigenvalue problems based on DG bilinear forms. We refer to [22] for some results along this direction.

4.3.3. Other coupling. We note that one can use other coupling mechanisms, such as mixed finite element methods [15], hybridized Galerkin [31, 32], and so on.

4.4. Handling nonlinearities. To handle nonlinear problems with GMsFEM, we use Discrete Empirical Interpolation Method (DEIM) and identify empirical modes and corresponding evaluations points (see [11]). Next, we briefly describe DEIM. Let $f(\nu) \in \mathbb{R}^n$ denote a nonlinear function where ν refers to any control parameter. We assume an approximation of the function f obtained by projecting it onto a subspace spanned by the basis functions (snapshots) $\Psi = (\psi_1, \cdots, \psi_m) \in \mathbb{R}^{n \times m}$ which are obtained by forward simulations. We write

$$f(\nu) \approx \Psi d(\nu). \tag{4.14}$$

To compute the coefficient vector d, we select m rows of (4.14) and invert a reduced system to compute $d(\nu)$. This can be formalized using the matrix P

$$\mathrm{P} = [e_{\wp_1}, \cdots, e_{\wp_m}] \in \mathbb{R}^{n\times m},$$

where $e_{\wp_i} = [0, \cdots, 0, 1, 0, \cdots, 0]^T \in \mathbb{R}^n$ is the $\wp_i^{\text{th}}$ column of the identity matrix $\mathrm{I}_n \in \mathbb{R}^{n\times n}$ for $i = 1, \cdots, m$. Multiplying Equation (4.14) by P^T and assuming that the matrix $\mathrm{P}^T\Psi$ is nonsingular, we obtain

$$f(\nu) \approx \Psi d(\nu) = \Psi(\mathrm{P}^T\Psi)^{-1}\mathrm{P}^T f(\nu). \tag{4.15}$$

To summarize, approximating the nonlinear function $f(\nu)$, as given by Equation (4.15), requires the following:

- computing the projection basis $\Psi = (\psi_1, \cdots, \psi_m)$.
- identifying the indices $\{\wp_1, \cdots, \wp_m\}$.

To determine the projection basis $\Psi = (\psi_1, \cdots, \psi_m)$, we collect function evaluations in an $n \times n_s$ matrix $\mathrm{F} = [f(\nu_1), \cdots, f(\nu_{n_s})]$ and employ Proper Orthogonal Decomposition (POD) to select the most energetic modes. This selection uses the eigenvalue decomposition of the square matrix $\mathrm{F}^T\mathrm{F}$ and form the important modes using the dominant eigenvalues. These modes are used as the projection basis in the approximation given by Equation (4.14). In Equation (4.15), the term $\Psi(\mathrm{P}^T\Psi)^{-1} \in \mathbb{R}^{n\times m}$ is computed once and stored. The $d(\nu)$ is computed using the values of the function $f(\nu)$ at m points with the indices $\wp_1, \cdots, \wp_m$ (identified using the DEIM algorithm). The resulting fewer evaluations of $f(\nu)$ yield significant computation savings.

We have considered the use of GMsFEM for nonlinear parabolic equations

$$\frac{\partial u}{\partial t} - \nabla\cdot(\kappa(x; u, \mu)\nabla u) = g(x) \quad \text{in } \Omega, \tag{4.16}$$

where we employed Newton method for the nonlinear solution strategy. When solving nonlinear PDEs, one writes the residual on the fine grid as

$$R(u) = 0, \tag{4.17}$$

where $R(u)$ is the residual of nonlinear PDE and u is the fine-grid solution. Here, both u and $R(u)$ are n-dimensional vectors defined on a fine grid. Using GMsFEM projection operator Φ, we project (4.17) onto the coarse degrees of freedom (noting that $u = \Phi z$ is an approximation of the fine-grid solution)

$$\Phi^T R(\Phi z) = 0. \tag{4.18}$$

This equation is formulated on the coarse degrees of freedom constructed on the coarse-grid; however, computing the residual $R(\Phi z)$ requires fine-grid evaluations. Moreover, computing the Jacobians in Newton iterations, defined as

$$J(z) = \nabla_z R(\Phi z),$$

also requires fine-grid evaluations. Here, our main goal is to use the multiscale DEIM to compute $R(\Phi z)$ and $J(z)$ efficiently. In particular, using the multiscale DEIM approximation, we can write

$$R(\Phi z) \approx \Psi d.$$

Consequently, the residual computation involves

$$\Phi^T \Psi d(z), \tag{4.19}$$

which can be efficiently computed by pre-computing $\Phi^T \Psi$. A similar procedure can be done for the Jacobian $J(z)$.

We detailed the multiscale Discrete Empirical Interpolation Methods (multiscale DEIM) in [11]. We stress the following main observations that are explored and ultimately are the motivations for the design of the multiscale DEIM procedure.

- In applications to multiscale PDEs, the nonlinear functional f needs to be evaluated at vectors u that are solutions obtained by reduced-order models. Thus, $f(\Phi z)$ needs to be computed in the span of coarse-grid snapshot vectors which has a reduced dimension.
- Due to the fact that multiscale basis functions are supported on a coarse-grid neighborhood, the DEIM approximation is obtained in each coarse-node neighborhood.
- More elaborate spectral selections are formulated to identify the elements of empirical interpolation vectors such that the resulting multiscale DEIM approximation is accurate in adequate norms that depend on physical parameters such as the contrast and small scales.

We refer to [7, 10] for more details and numerical results.

4.5. Applications. The applications of GMsFEM in various fields are studied in the literature. Below, we briefly describe some of these applications.

- *Compressible flow.*

$$\frac{\partial p}{\partial t} - \operatorname{div}\big(\kappa(x)\,\nabla p\big) = q, \tag{4.20}$$

  where $p(x,t)$ denotes the time-varying pressure within a specified domain D. Here, $\kappa(x)$ is a heterogeneous permeability field. We refer to [20].

- *Wave equation.*

$$\frac{\partial^2}{\partial t^2} p - \kappa \nabla \cdot (\rho^{-1} \nabla p) = q, \tag{4.21}$$

  where $p = p(x,t)$ is the pressure wavefield, $\kappa = \kappa(x)$ is the bulk modulus of the media which may vary greatly below the dominant seismic wavelength, $\rho = \rho(x)$ is the density of medium, and $q = q(x,t)$ is the external force term. Assuming constant κ and normalizing the density ρ with it, so that (4.21) may be written as

$$\frac{\partial^2}{\partial t^2} p - \nabla \cdot (c^2 \nabla p) = q, \tag{4.22}$$

  where $c^2 = \kappa/\rho$. We refer to [17, 35] for the development of continuous and discontinuous Galerkin methods and their applications to seismic wave propagation.

- *Elasticity equations.*

$$\begin{aligned} -\operatorname{div}\sigma(u) &= f \\ \sigma(u) &= C : e(u) \end{aligned} \tag{4.23}$$

where σ is the stress tensor, e is defined as $e = e(u) = \frac{1}{2}(\nabla u + \nabla u^T)$, $C = C(x), x \in \Omega$ is the fourth order elasticity multiscale tensor, and u is the displacement field. We refer to [14, 34] for further discussions for static and elastic wave equations.

- *Brinkman equation.*

$$\begin{aligned} \nabla p(x) - \mu \Delta u(x) + \kappa^{-1} u(x) &= f(x) \\ \text{div}(u(x)) &= 0. \end{aligned} \tag{4.24}$$

The application of GMsFEM to Brinkman equation is studied in [33].

- *Two-phase incompressible flow and transport.*

Multi-phase fluid flow is another area where GMsFEM may be used as an effective solution technique. For this selected application we consider a heterogeneous oil reservoir which is confined to a global domain D. We consider an immiscible two-phase system containing water and oil (where the respective subscripts w and o are often used) that is incompressible. We also assume a gravity-free environment and that the pore space is fully saturated. Then, a statement of conservation of mass combined with Darcy's law allows us to write the governing equations of the flow as

$$\text{div}(v) = q, \quad \text{where} \quad v = -\lambda(S)\kappa(x)\nabla p, \tag{4.25}$$

and

$$\frac{\partial S}{\partial t} + \text{div}(f(S)v) = q_w, \tag{4.26}$$

where $p(x,t)$ denotes the pressure, $v(x,t)$ is the Darcy velocity, $S(x,t)$ is the water saturation, and $\kappa(x)$ is the high-contrast permeability coefficient. The total mobility $\lambda(S)$ and the flux function $f(S)$ are respectively given by

$$\lambda(S) = \frac{\kappa_{rw}(S)}{\mu_w} + \frac{\kappa_{ro}(S)}{\mu_o}, \quad \text{and} \quad f(S) \quad = \frac{\kappa_{rw}(S)/\mu_w}{\lambda(S)}, \tag{4.27}$$

where κ_{rj} $(j = w, o)$ is the relative permeability of the phase j, and μ_j $(j = w, o)$ is the respective fluid viscosity. We refer to [9] for more details on applications of GMsFEM to two-phase flow models.

- *Monotone nonlinear operators.*

$$\text{div}\kappa(x, \nabla u) = f.$$

We have studied the case $\kappa(x, \nabla u) = \kappa(x)|\nabla u|^{\nu-2}\nabla u$ in [24].

- *Uncertainty quantification in flow.*

We have developed multi-level Monte Carlo and multi-level Markov chain Monte Carlo using GMsFEM framework for flow problems in [30]

4.6. Adaptivity. The adaptivity in GMsFEM is studied in [16] where we derive an *a-posteriori* error indicator for the Generalized Multiscale Finite Element Method (GMsFEM) framework. This error indicator is further used to develop an adaptive enrichment algorithm for the linear elliptic equation with multiscale high-contrast coefficients. We consider two kinds of error indicators where one is based on the L^2-norm of the local residual and the

other is based on the weighted H^{-1}-norm of the local residual where the weight is related to the coefficient of the elliptic equation. We show that the use of weighted H^{-1}-norm residual gives a more robust error indicator which works well for cases with high contrast media. The convergence analysis of the method is given. Numerical results are presented that demonstrate the robustness of the proposed error indicators.

4.7. Global-local model reduction techniques. We have developed global-local model reduction techniques that use GMsFEM to speed-up global model reduction techniques. In these techniques, the main idea is to solve for global snapshots using adaptive multiscale methods and perform model reduction for a global problem using solutions on a coarse grid. Some results for linear and nonlinear problems can be found in [7, 20, 36].

5. Conclusions

In this paper, we discuss multiscale model reduction through the use of the Generalized Multiscale Finite Element Method (GMsFEM). We outline the basic concepts associated with the systematic enrichment of coarse solution spaces, and describe the offline-online procedure that is used in the construction of multiscale basis functions. We discuss various applications. For further details regarding each application, we direct the interested reader to pertinent references.

Acknowledgments. YE is grateful to DOE for the support. Y. Efendiev's work is partially supported by the DOE. I am grateful to my collaborators with whom I worked on developing multiscale methods.

References

[1] J.E. Aarnes, *On the use of a mixed multiscale finite element method for greater flexibility and increased speed or improved accuracy in reservoir simulation*, SIAM J. Multiscale Modeling and Simulation **2** (2004), 421–439.

[2] J.E. Aarnes and Y. Efendiev, *Mixed multiscale finite element for stochastic porous media flows*, SIAM J. Sci. Comput. **30 (5)** (2008), 2319–2339.

[3] J.E. Aarnes, Y. Efendiev, and L. Jiang, *Analysis of multiscale finite element methods using global information for two-phase flow simulations*, SIAM J. Multiscale Modeling and Simulation **7** (2008), no. 2, 655–676.

[4] J.E. Aarnes and T. Hou, *Multiscale domain decomposition methods for elliptic problems with high aspect ratios*, Acta Math. Appl. Sin. Engl. Ser. **18** (2002), 63–76.

[5] J.E. Aarnes, S. Krogstad, and K.-A. Lie, *A hierarchical multiscale method for two-phase flow based upon mixed finite elements and nonuniform grids*, SIAM J. Multiscale Modeling and Simulation **5** (2006), no. 2, 337–363.

[6] A. Abdulle, *Multiscale method based on discontinuous galerkin methods for homogenization problems*, C.R. Math. Acad. Sci. Paris **346** (2008), no. 1-2, 97–102.

[7] M. Alatoibi, V. Calo, Y. Efendiev, J. Galvis, and M. Ghommem, *Global-local nonlinear model reduction for flows in heterogeneous porous media*, CMAME (2014), submitted.

[8] T. Arbogast, *Implementation of a locally conservative numerical subgrid upscaling scheme for two-phase Darcy flow*, Comput. Geosci. **6** (2002), 453–481.

[9] L. Bush, V. Ginting, and M. Presho, *Application of a conservative, generalized multiscale finite element method to flow models*, J. Comput. Appl. Math. **260** (2014), 395–409.

[10] V. Calo, Y. Efendiev, J. Galvis, and M. Ghommem, *Multiscale empirical interpolation for solving nonlinear pdes using generalized multiscale finite element methods*, JCP (2014), submitted.

[11] S. Chaturantabut and D. C. Sorensen, *Discrete empirical interpolation for nonlinear model reduction*, SIAM J. Sci. Comput. **32** (2010), no. 5, 2737–2764.

[12] Y. Chen and L. Durlofsky, *An ensemble level upscaling approach for efficient estimation of fine-scale production statistics using coarse-scale simulations*, SPE Reservoir Simulation Symposium (Houston, Texas, U.S.A.), no. 106086-MS, Society of Petroleum Engineers, 2 2007.

[13] Y. Chen, L. Durlofsky, M. Gerritsen, and X. Wen, *A coupled local-global upscaling approach for simulating flow in highly heterogeneous formations*, Advances in Water Resources **26** (2003), 1041–1060.

[14] E. Chung, Y. Efendiev, and S. Fu, *Generalized multiscale finite element methods for elasticity equations in heterogeneous media*, in preparation.

[15] E. Chung, Y. Efendiev, and C. S. Lee, *Generalized mixed multiscale finite element method for flows in heterogeneous media*, (2014), submitted.

[16] E. Chung, Y. Efendiev, and G. Li, *An adaptive Generalized Multiscale Finite Element Method for high-contrast flow problems*, to appear in JCP.

[17] E. Chung, Y. Efendiev, and W. Tat, *Generalized Multiscale Finite Element Method for wave propagation in heterogeneous media*, SIAM MMS (2014), submitted.

[18] L.J. Durlofsky, *Numerical calculation of equivalent grid block permeability tensors for heterogeneous porous media*, Water Resour. Res. **27** (1991), 699–708.

[19] W. E and B. Engquist, *Heterogeneous multiscale methods*, Comm. Math. Sci. **1** (2003), no. 1, 87–132.

[20] Y. Efendiev, J. Galvis, and E. Gildin, *Local-global multiscale model reduction for flows in highly heterogeneous media*, JCP **231** (2012), 8100–8113.

[21] Y. Efendiev, J. Galvis, and T. Hou, *Generalized multiscale finite element methods*, JCP **251** (2013), 116–135.

[22] Y. Efendiev, J. Galvis, R. Lazarov, M. Moon, and M. Sarkis, *Generalized multiscale finite element method. symmetric interior penalty coupling*, Journal of Computational Physics **255** (2013), no. 0, 1 – 15.

[23] Y. Efendiev, J. Galvis, G. Li, and M. Presho, *Generalized multiscale finite element methods. oversampling strategies*, IJMM, to appear.

[24] Y. Efendiev, J. Galvis, M. Presho, and J. Zhou, *A multiscale enrichment procedure for nonlinear monotone operators*, Math. Model. Numer. Anal. (M2AN) (2014), to appear.

[25] Y. Efendiev, J. Galvis, and X.H. Wu, *Multiscale finite element methods for high-contrast problems using local spectral basis functions*, Journal of Computational Physics **230** (2011), 937–955.

[26] Y. Efendiev, V. Ginting, T. Hou, and R. Ewing, *Accurate multiscale finite element methods for two-phase flow simulations*, Journal of Computational Physics **220** (2006), 155–174.

[27] Y. Efendiev and T. Hou, *Multiscale Finite Element Methods: Theory and Applications*, Surveys and Tutorials in the Applied Mathematical Sciences, vol. 4, Springer, New York, 2009.

[28] Y. Efendiev, T. Hou, and X.H. Wu, *Convergence of a nonconforming multiscale finite element method*, SIAM J. Numer. Anal. **37** (2000), 888–910.

[29] Y. Efendiev, T. How, and V. Ginting, *Multiscale finite element methods for nonlinear problems and their applications*, Comm. Math. Sci. **2** (2004), 553–589.

[30] Y. Efendiev, B. Jin, M. Presho, and X. Tan, *Multilevel markov chain monte carlo method for high-contrast single-phase flow problems*, CiCP (2014), submitted.

[31] Y. Efendiev, R. Lazarov, M. Moon, and K. Shi, *A spectral multiscale hybridizable discontinuous galerkin method for second order elliptic problems*, CMAME (2014), submitted.

[32] Y. Efendiev, R. Lazarov, and K. Shi, *A multiscale HDG method for second order elliptic equations. Part I. Polynomial and homogenization-based multiscale spaces*, ArXiv e-prints, arXiv:1310.2827 (2013).

[33] J. Galvis, S. Ke, and G. Li, *Multiscale model reduction for brinkman's flows in heterogeneous media using generalized multiscale finite element method*, JCAM (2014), submitted.

[34] K. Gao, S. Fu, R. Gibson, E. Chung, and Y. Efendiev, *Generalized multiscale finite element method for elastic wave equations*, Expanded SEG Abstracts 2014, submitted.

[35] K. Gao, R. Gibson, E. Chung, Y. Efendiev, and S. Fu, *A multiscale method for elastic wave equation modeling*, Expanded SEG Abstract (2013).

[36] M. Ghommem, V. M. Calo, and Y. Efendiev, *Mode decomposition methods for flows in high-contrast porous media: Part I. Global approach (Part II. Global-local approach)*, Journal of Computational Physics **253 (257)** (2013), 226–238 (400–413).

[37] T. Hou and X.H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys. **134** (1997), 169–189.

[38] T. Hughes, G. Feijoo, L. Mazzei, and J. Quincy, *The variational multiscale method - a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg. **166** (1998), 3–24.

[39] P. Jenny, S.H. Lee, and H. Tchelepi, *Multi-scale finite volume method for elliptic problems in subsurface flow simulation*, J. Comput. Phys. **187** (2003), 47–67.

[40] V. Jikov, S. Kozlov, and O. Oleinik, *Homogenization of differential operators and integral functionals*, Springer-Verlag, Translated from Russian, 1994.

[41] A. Matache and C. Schwab, *Homogenization via p-fem for problems with microstructure*, Appl. Numer. Math. **33** (2000), 43–59.

[42] M. Ohlberger, *A posteriori error estimates for the heterogeneous multiscale finite element method for elliptic homogenization problems*, SIAM J. Multiscale Modeling and Simulation **4** (2005), no. 1, 88–114.

[43] H. Owhadi and L. Zhang, *Metric-based upscaling*, Comm. Pure. Appl. Math. **60** (2007), 675–723.

[44] X.H. Wu, Y. Efendiev, and T.Y. Hou, *Analysis of upscaling absolute permeability*, Discrete and Continuous Dynamical Systems, Series B. **2** (2002), 158–204.

Department of Mathematics, Texas A & M University, College Station, TX; Numerical Porous Media SRI Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

E-mail: efendiev@math.tamu.edu

Discontinuous Galerkin method for time-dependent convection dominated partial differential equations

Chi-Wang Shu

Abstract. In this lecture we give an introduction to discontinuous Galerkin (DG) methods for solving time-dependent convection dominated partial differential equations (PDEs). DG methods form a class of finite element methods. Differently from classical finite element methods, which are built upon spaces containing continuous, piecewise polynomial functions, DG methods are built upon function spaces containing piecewise polynomials (or other simple functions) which are allowed to be completely discontinuous across element interfaces. Using finite element terminologies, DG methods are the most extreme case of *nonconforming* finite element methods. DG methods are most natural and most successful for solving hyperbolic conservation laws which have generic discontinuous solutions. Moreover, in recent years stable and convergent DG methods have also been designed for convection dominated PDEs containing higher order spatial derivatives, such as convection diffusion equations and KdV equations. We will emphasize the guiding principles for the design and analysis, and recent development and applications of the DG methods for solving time-dependent convection dominated PDEs.

Mathematics Subject Classification (2010). Primary 65M60, 65M20, 65M12, 65M15.

Keywords. Discontinuous Galerkin method, time-dependent convection dominated partial differential equations, hyperbolic equations, convection-diffusion equations, stability, error estimates, superconvergence, limiters.

1. Introduction

Discontinuous Galerkin (DG) methods form a class of finite element methods. Differently from classical finite element methods, which are built upon spaces containing continuous, piecewise polynomial functions, DG methods are built upon function spaces containing piecewise polynomials (or other simple functions) which are allowed to be completely discontinuous across element interfaces. Using finite element terminologies, DG methods are the most extreme case of *nonconforming* finite element methods. In this lecture we concentrate on DG methods for time-dependent, convection dominated partial differential equations (PDEs).

The earliest DG method was introduced in 1973 by Reed and Hill in a Los Alamos technical report [65], in which the equations for neutron transport, which are time-independent linear hyperbolic equations, were solved. A major development of DG methods was carried out in a series of papers [15, 17, 19–21], in which the authors established a framework to easily solve *nonlinear* time-dependent hyperbolic equations, such as the Euler equations

Proceedings of the International Congress of Mathematicians, Seoul, 2014

of compressible gas dynamics. The DG methods in [15, 17, 19–21] belong to the class of method-of-lines, namely the DG discretization is used only for the spatial variables, and explicit, nonlinearly stable high order Runge-Kutta methods [73] are used to discretize the time variable. Important features of the DG methods in [15, 17, 19–21] include the usage of finite volume methodologies, such as exact or approximate Riemann solvers as interface fluxes and total variation bounded (TVB) nonlinear limiters [71] to control spurious oscillations in the presence of strong shocks.

In recent years there has been an explosion of activities related to the development, analysis and applications of DG methods. Among the areas of applications we could mention aeroacoustics, electro-magnetism, gas dynamics, granular flows, magneto-hydrodynamics, meteorology, modeling of shallow water, oceanography, oil recovery simulation, semiconductor device simulation, transport of contaminant in porous media, turbomachinery, turbulent flows, viscoelastic flows and weather forecasting. For earlier work on DG methods, we refer to the survey paper [16], and other papers in that Springer volume, which contains the conference proceedings of the First International Symposium on Discontinuous Galerkin Methods held at Newport, Rhode Island in 1999. The lecture notes [13] is a good reference for many details, as well as the extensive review paper [23]. The review paper [91] covers the local DG method for PDEs containing higher order spatial derivatives. There are three recent special journal issues devoted to the DG method [24, 25, 27], which contain many interesting papers on DG method in all aspects including algorithm design, analysis, implementation and applications. There are also a few recent books and lecture notes [32, 43, 49, 67, 72] on DG methods.

2. DG method for hyperbolic conservation laws

DG methods are most successful for solving hyperbolic conservation laws. As mentioned in the previous section, the first DG method [65] was designed to solve steady state hyperbolic conservation laws. We concentrate on time-dependent PDEs in this lecture. In one-space dimension, a hyperbolic conservation law is given by

$$u_t + f(u)_x = 0. \tag{2.1}$$

In the system case u is a vector, then the Jacobian matrix $f'(u)$ is required to be diagonalizable with real eigenvalues. In two-space dimensions, the equation is

$$u_t + f(u)_x + g(u)_y = 0.$$

Important properties for the solutions of hyperbolic conservation laws include:

- The solution u may become discontinuous regardless of the smoothness of the initial condition. Therefore, we must consider weak solutions instead of classical strong solutions.
- Weak solutions may not be unique. The unique, physically relevant weak solution, also referred to as the entropy solution, satisfies additional entropy inequalities

  $$U(u)_t + F(u)_x \le 0 \tag{2.2}$$

  in the distribution sense, where $U(u)$ is a convex scalar function of u and the entropy flux $F(u)$ satisfies $F'(u) = U'(u)f'(u)$.

For more properties of entropy solutions of hyperbolic conservation laws, we refer to [74].

The starting point for the design of DG methods can be described as follows. Suppose we are solving the equation (2.1) over the interval [0,1], with periodic boundary condition for simplicity, then we first divide [0,1] into N cells

$$0 = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N+\frac{1}{2}} = 1,$$

and denote

$$I_j = \left(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\right), \quad x_j = \frac{1}{2}\left(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}}\right), \quad h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$$

as the cells, cell centers and cell lengths respectively. We also define $h = h_{\max} = \max_j h_j$ and $h_{\min} = \min_j h_j$, and we consider only regular meshes, that is $h_{\max} \leq \lambda h_{\min}$ where $\lambda \geq 1$ is a constant during mesh refinement. If $\lambda = 1$, then the mesh is uniformly distributed. If we multiply the equation (2.1) with an arbitrary smooth test function v, integrate over the cell $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, and integrate by parts, we obtain

$$\int_{I_j} u_t v dx - \int_{I_j} f(u) v_x dx + f(u_{j+\frac{1}{2}}) v_{j+\frac{1}{2}} - f(u_{j-\frac{1}{2}}) v_{j-\frac{1}{2}} = 0. \tag{2.3}$$

Here, we have used the short notation $u_{j+\frac{1}{2}} = u(x_{j+\frac{1}{2}}, t)$ etc. Notice that (2.3) is not a scheme yet, rather it is an equality satisfied by the exact solution u of (2.1) and any smooth test function v. We now attempt to convert it to a numerical scheme. For this purpose we define the DG finite element space as

$$V_h^k = \{v : v|_{I_j} \in \mathcal{P}^k(I_j), j = 1, \cdots, N\}, \tag{2.4}$$

where $\mathcal{P}^k(I_j)$ denotes the space of polynomials in I_j of degree at most k. This polynomial degree k can actually change from cell to cell, but we assume it is a constant in this lecture for simplicity. We now attempt to replace u and v in the equality (2.3) by u_h and v_h, both of them taken from the DG space V_h^k. However, the intercell boundary terms $f(u_{j+\frac{1}{2}})$, $v_{j+\frac{1}{2}}$ etc. are not well defined when u and v are replaced by u_h and v_h in V_h^k, since in this space the functions are *discontinuous* at the cell interfaces. This is an inconvenience but also an opportunity for the design of DG methods. A good choice to resolve these ambiguities leads to good DG schemes which are stable and optimal rate accurate. It is also here that one of the important concepts from finite volume schemes is borrowed, namely monotone numerical fluxes (for the scalar case), or exact or approximate Riemann solvers (for the system case). We refer to [47] for more details. Thus we take

- a single valued monotone numerical flux to replace $f(u_{j+\frac{1}{2}})$:

$$\hat{f}_{j+\frac{1}{2}} = \hat{f}((u_h)^-_{j+\frac{1}{2}}, (u_h)^+_{j+\frac{1}{2}}) \tag{2.5}$$

  where the numerical flux $\hat{f}$ satisfies consistency $\hat{f}(u,u) = f(u)$; monotonicity $\hat{f}(\uparrow,\downarrow)$ (non-decreasing in the first argument and non-increasing in the second argument, for the scalar case only), and Lipschitz continuity with respect to both arguments;

- values from inside I_j for the test function v_h

$$(v_h)^-_{j+\frac{1}{2}}, \qquad (v_h)^+_{j-\frac{1}{2}}.$$

Hence the DG scheme for solving (2.1) is: find the unique function $u_h = u_h(\cdot,t) \in V_h^k$ such that, for all test functions $v_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\int_{I_j} (u_h)_t \, v_h dx - \int_{I_j} f(u_h)(v_h)_x dx + \hat{f}_{j+\frac{1}{2}}(v_h)^-_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}(v_h)^+_{j-\frac{1}{2}} = 0 \tag{2.6}$$

where the numerical flux $\hat{f}_{j+\frac{1}{2}}$ is defined by (2.5).

The semi-discrete DG scheme (2.6) can be discretized in time by the total variation diminishing (TVD) Runge-Kutta methods [73], also referred to in later literature as the strong-stability-preserving (SSP) time discretizations [29, 30]. For the semi-discrete scheme:

$$\frac{du}{dt} = L(u)$$

where $L(u)$ is a discretization of the spatial operator, the third order TVD Runge-Kutta method in [73] is simply:

$$\begin{aligned} u^{(1)} &= u^n + \Delta t L(u^n) \\ u^{(2)} &= \frac{3}{4}u^n + \frac{1}{4}u^{(1)} + \frac{1}{4}\Delta t L(u^{(1)}) \\ u^{n+1} &= \frac{1}{3}u^n + \frac{2}{3}u^{(2)} + \frac{2}{3}\Delta t L(u^{(2)}). \end{aligned} \tag{2.7}$$

We now briefly discuss several important properties and advantages of the DG method for solving hyperbolic conservation laws.

- The DG method allows an easy handling of complicated geometry and boundary conditions. Arbitrary triangulations allow an easy fit to general geometry, and the local nature of the method (communication with the neighbors only through the numerical fluxes) allows an easy implementation of commonly used boundary conditions. While these advantages are mostly common to all finite element methods, the DG method has the additional advantage of allowing "hanging nodes" naturally, which are also referred to as non-conforming meshes in finite elements.
- The DG method has a compact stencil. Communication is needed only with immediate neighbors through numerical fluxes on cell interfaces, regardless of the order of the scheme. In comparison, high order finite difference or finite volume schemes would need a wide stencil in order to obtain high order polynomial interpolations or reconstructions.
- The DG method is explicit. Because of the discontinuous basis, the mass matrix is local to the cell, resulting in an explicit time stepping for which no large linear system needs to be solved.
- The DG method has excellent parallel efficiency, largely because of its compact stencil, minimum communication with neighbors and explicit time discretization. It can achieve up to 99% parallel efficiency for static mesh and over 80% parallel efficiency for dynamic load balancing with adaptive meshes, see, e.g. [4, 66]. The method is also friendly to the GPU environment and can achieve amazing speedup there [44].

- The DG method is one of the very few high order numerical methods for which a cell entropy inequality for the square entropy and consequently an L^2 stability can be proved, for arbitrary nonlinear equations in any spatial dimension and any triangulations, for any polynomial degrees, without limiters or assumption on solution regularity. For the entropy $U(u) = \frac{u^2}{2}$, we can find a consistent entropy flux $\hat{F}_{j+1/2}$ such that
$$\frac{d}{dt}\int_{I_j} U(u_h)dx + \hat{F}_{j+1/2} - \hat{F}_{j-1/2} \leq 0, \tag{2.8}$$
which is a direct approximation to the analytic entropy inequality (2.2) in the cell I_j, hence we refer to it as the cell entropy inequality. Summing the cell entropy inequality (2.8) over j, we obtain
$$\frac{d}{dt}\int_a^b (u_h)^2 dx \leq 0,$$
which establishes the L^2 stability of the DG solution. The proof of the cell entropy inequality and the associated L^2 stability of the DG method for scalar equations is given in [40]. The conclusion also holds for symmetric hyperbolic systems as shown in [33]. The cell entropy inequality also holds for fully discrete Runge-Kutta DG (RKDG) methods with the third order TVD Runge-Kutta time discretization (2.7), at least for the linear equations [101].
- The DG method can be proved to converge in the optimal rate of $(k+1)$-th order accuracy or at least in the rate of $(k+\frac{1}{2})$-th order accuracy in the L^2-norm, for smooth solutions when piecewise polynomials of degree k are used, regardless of the structure of the meshes. The earlier work in such error estimates include those in [42, 46], for linear, steady state hyperbolic equations. Error estimates for fully discretized RKDG schemes for linear and nonlinear scalar conservation laws and symmetrizable hyperbolic systems can be found in [56, 57, 99–101].
- The DG method has excellent superconvergence properties for linear and nonlinear hyperbolic equations. It is proved in [18] that the DG solution is $(2k+1)$-th order superconvergent in the negative norm for general meshes, and a post-processed solution based on convolution of the DG solution with a locally defined kernel [5] is $(2k+1)$-th order superconvergent in the strong L^2-norm on translation invariant meshes, for smooth solutions when piecewise polynomials of degree k are used. These results have been generalized to one-sided post-processing near the boundaries [68], structured triangular meshes [59], non-uniform meshes [26], and nonlinear problems [38]. It has also been applied to aeroacoustics [69] and computer graphics [75].
- The DG solution has been proved to be $(k+3/2)$-th or $(k+2)$-th order superconvergent to a special projection of the exact solution, as a consequence the error does not grow in time up to $t = O(\frac{1}{\sqrt{h}})$ or $t = O(\frac{1}{h})$, for both linear and nonlinear hyperbolic equations, for smooth solutions when piecewise polynomials of degree k are used. In [10], Cheng and Shu started this line of study by obtaining $(k+3/2)$-th order superconvergence for linear, time-dependent hyperbolic equations in one-dimension, with uniform meshes and periodic boundary conditions. The proof is based on Fourier analysis and is carried out only for the piecewise linear $k=1$ case, however numerical results confirm the validity for higher k's. Another important consequence of this superconvergence result is that the constant C in front of the $h^{k+3/2}$ error term only

grows at most linearly with time t, therefore the standard L^2 error does not grow for a very long time $t \sim 1/\sqrt{h}$. This analysis verifies an observation by practitioners, that the error of the DG solution for wave propagation does not seem to grow much with time. The result in [10] is improved in [12] to general polynomial degree k, on non-uniform regular meshes, and without periodic boundary conditions. The technique used in [12] is a finite element type, not a Fourier analysis. In [96], the result in [12] is further improved to $(k+2)$-th order superconvergence. This half-order increase in the analysis is highly non-trivial and involves subtle handling of cancellation of errors during time evolution. The result in [96] is optimal. In [58], $(k+3/2)$-th order superconvergence is proved for scalar *nonlinear* conservation laws with a fixed wind direction in one space dimension.

- Besides error estimates for smooth solutions, it is perhaps more relevant to study error estimates for discontinuous or otherwise singular solutions, which are generic for hyperbolic partial differential equations. Optimal L^2 error estimates of the DG method for discontinuous solutions of linear hyperbolic equations in a region $O(\sqrt{h}\log h)$ away from the discontinuities are proved in [14] for piecewise linear DG method on uniform meshes, and in [102] for general RKDG methods with third order TVD Runge-Kutta time discretization.
- Bound preserving limiters, which can preserve strict maximum principle for scalar hyperbolic equations and positivity of relevant physical quantities for hyperbolic systems (e.g. density and pressure for Euler systems for gas dynamics and water height for shallow water equations) while maintaining the original high order accuracy of the DG schemes, have been designed in a series of recent papers [76, 81, 103–107]. These limiters have significantly improved the robustness of DG solutions while maintaining their originally designed high order accuracy.
- Even though the DG schemes for conservation laws are L^2 stable, for solving problems with strong discontinuities, the DG solution may still generate spurious numerical oscillations. In practice, especially for nonlinear problems containing strong shocks, we often need to apply nonlinear limiters to control these oscillations. Most of the limiters studied in the literature come from the methodologies of finite volume high resolution schemes. Earlier limiters include the TVD and TVB limiters [31, 71], and the moment-based limiters [4, 6]. More recently, limiters based on the weighted essentially non-oscillatory (WENO) methodology are designed with the objective of maintaining the high order accuracy even if they take effect in smooth cells. These limiters are based on the WENO methodology for finite volume schemes [41, 52], and involve nonlinear reconstructions of the polynomials in troubled cells using the information of neighboring cells. The WENO reconstructed polynomials have the same high order of accuracy as the original polynomials when the solution is smooth, and they are (essentially) non-oscillatory near discontinuities. Qiu and Shu [63] and Zhu et al. [111] designed WENO limiters using the usual WENO reconstruction based on cell averages of neighboring cells as in [34, 41, 70]. This limiter needs to use the information from not only the immediate neighboring cells but also neighbors' neighbors, making it complicated to implement in multi-dimensions, especially for unstructured meshes [34, 109, 111]. It also destroys the local data structure of the base DG scheme (which needs only to communicate with immediate neighbors). The effort in [61, 62] attempts to construct Hermite type WENO approximations, which use the informa-

tion of not only the cell averages but also the lower order moments such as slopes, to reduce the spread of reconstruction stencils. However for higher order methods the information of neighbors' neighbors is still needed. More recently, Zhong and Shu [110] developed a new WENO limiting procedure for RKDG methods on structured meshes. The main advantage of this limiter is its simplicity in implementation, as it uses only the information from immediate neighbors and the linear weights are always positive. This simplicity is more prominent for multi-dimensional unstructured meshes, which is studied in [112] for two-dimensional unstructured triangular meshes. The WENO limiters are typically applied only in designated "troubled cells", in order to save computational cost and to minimize the influence of accuracy in smooth regions. Therefore, a troubled cell indicator is needed, to correctly identify cells near discontinuities in which the limiters should be applied. Qiu and Shu in [64] have compared several troubled cell indicators. In practice, the TVB indicator [71] and the KXRCF indicator [45] are often the best choices.

- Because of the local nature and discontinuous basis functions, DG methods are extremely flexibility for both h (refining meshes) and p (adjusting polynomial degrees in different cells) adaptivity. An example of the application of such adaptivity can be found in [66].

3. DG method for convection diffusion equations

While the DG method is most natural and highly successful for solving hyperbolic equations which have generic discontinuous solutions, in applications one often encounters convection dominated PDEs which contain higher order spatial derivatives. A typical example would be a convection dominated convection diffusion equation, for example the compressible Navier-Stokes equations in gas dynamics with high Reynolds numbers. It would be desirable to have a DG method which is stable and accurate for such equations.

Let us look at the simple heat equation

$$u_t - u_{xx} = 0 \tag{3.1}$$

as an example. A straightforward generalization of the DG method from the hyperbolic equation (2.1) is to write down the same scheme (2.6) and replace $f(u)$ by $-u_x$ everywhere: find $u_h \in V_h^k$ such that, for all test functions $v_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\int_{I_j} (u_h)_t v_h dx + \int_{I_j} (u_h)_x (v_h)_x dx - \widehat{u_x}_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} + \widehat{u_x}_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = 0. \tag{3.2}$$

Of course, we still need to define the numerical flux $\widehat{u_x}_{j+\frac{1}{2}}$. Lacking an upwinding consideration for the choice of this numerical flux and considering that diffusion is isotropic, a natural choice for the flux could be the central flux

$$\widehat{u_x}_{j+\frac{1}{2}} = \frac{1}{2}\left(((u_h)_x)^-_{j+\frac{1}{2}} + ((u_h)_x)^+_{j+\frac{1}{2}} \right). \tag{3.3}$$

However, numerical experiments show that the scheme (3.2) with the numerical flux (3.3) is terrible! The errors do not decay with mesh refinement, and the numerical solution, although

seemingly convergent with mesh refinement, does not converge to the correct solution of the PDE with the given initial condition.

It is proven in [98] that this "bad" DG method for the heat equation is actually consistent with the heat equation (3.2) but is (very weakly) unstable.

This "bad" DG scheme reminds us that we have to be cautious in designing DG schemes for solving PDEs containing higher than first order spatial derivatives, such as the heat equation (3.1). A "good" DG method for the heat equation (3.1) is the local DG (LDG) method [2, 22]. First, we rewrite the heat equation (3.1) as

$$u_t - q_x = 0, \qquad q - u_x = 0, \tag{3.4}$$

and *formally* write out the DG scheme as: find $u_h, q_h \in V_h^k$ such that, for all test functions $v_h, w_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\begin{aligned} &\int_{I_j} (u_h)_t v_h dx + \int_{I_j} q_h (v_h)_x dx - \hat{q}_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} + \hat{q}_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = 0 \\ &\int_{I_j} q_h w_h dx + \int_{I_j} u_h (w_h)_x dx - \hat{u}_{j+\frac{1}{2}} (w_h)^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} (w_h)^+_{j-\frac{1}{2}} = 0. \end{aligned} \tag{3.5}$$

Notice that, by the second equality in (3.5), q_h can be locally (within the cell I_j) solved and eliminated, hence the method is referred to as a *local* DG method.

A key ingredient in the design of the LDG method is the choice of the numerical fluxes $\hat{u}$ and $\hat{q}$ (remember: no upwinding principle exists for a guidance). The best choice for the numerical fluxes is the following alternating flux

$$\hat{u}_{j+\frac{1}{2}} = (u_h)^-_{j+\frac{1}{2}}, \qquad \hat{q}_{j+\frac{1}{2}} = (q_h)^+_{j+\frac{1}{2}}. \tag{3.6}$$

The other way around also works

$$\hat{u}_{j+\frac{1}{2}} = (u_h)^+_{j+\frac{1}{2}}, \qquad \hat{q}_{j+\frac{1}{2}} = (q_h)^-_{j+\frac{1}{2}}.$$

With such choice of numerical fluxes, the scheme (3.5) is L^2 stable and has optimal convergence of $O(h^{k+1})$ in the L^2 norm for P^k elements [22, 72].

The conclusions are valid for general nonlinear multi-dimensional convection diffusion equations

$$u_t + \sum_{i=1}^{d} f_i(u)_{x_i} - \sum_{i=1}^{d} \sum_{j=1}^{d} (a_{ij}(u) u_{x_j})_{x_i} = 0, \tag{3.7}$$

where $a_{ij}(u)$ are entries of a symmetric and semi-positive definite matrix. LDG methods which are L^2 stable and convergent can be obtained, see [22, 86].

Regarding superconvergence, similar results as those for hyperbolic equations are available for the LDG schemes solving convection-diffusion equations, either in the negative norm [39] or in the error to a special projection of the exact solution [11, 12, 97].

Maximum-principle preserving uniformly second order (P^1) LDG method for nonlinear convection-diffusion equations including two-dimensional incompressible Navier-Stokes equations in vorticity-streamfunction formulation, for arbitrary two-dimensional regular triangulations without acute-angle restrictions, has been obtained in [108].

One major advantage of LDG method for convection-diffusion equations is that it works well for convection-dominated situation with small or even locally vanishing diffusion coefficients. A typical example is the porous medium equation

$$u_t = \Delta(u^m), \quad m > 1.$$

The solution to this PDE may contain singularities (discontinuities in the first derivative) which has a finite propagation speed, similar to hyperbolic conservation laws. Negative density u leads to ill-posedness and instability of the code. Our maximum-principle preserving DG scheme however works well, see [108].

Besides LDG methods, there are also a few other types of DG methods for convection-diffusion equations:

- Internal penalty DG methods, including the symmetric internal penalty DG (SIPG) method [1, 77] and the non-symmetric internal penalty DG (NIPG) method [3, 60]. A penalty parameter is involved which should be chosen in suitable ranges. There are other types of DG methods involving the internal penalty methodology, for example the direct discontinuous Galerkin (DDG) methods [50, 51].
- Ultra weak DG methods, which is based on integration by parts twice to put all derivatives on test functions, and then introducing numerical fluxes for both the function and its first derivative. A penalty term is still needed. See [9].

4. DG method for higher order convection dominated PDEs

DG methods can be designed for higher (than second) order PDEs. We will concentrate on LDG methods and will discuss dispersive wave equations (usually odd order) and diffusive equations (usually even order) separately below.

4.1. LDG method for dispersive wave equations. Let us look at the Korteweg-de Vries (KdV) equation:

$$u_t + (\alpha u + \beta u^2)_x + \sigma u_{xxx} = 0.$$

More generally, we can look at the fully nonlinear version in one-dimension

$$u_t + f(u)_x + (r'(u)g(r(u)_x)_x)_x = 0$$

and in multi-dimensions

$$u_t + \sum_{i=1}^{d} f_i(u)_{x_i} + \sum_{i=1}^{d} \left(r_i'(u) \sum_{j=1}^{d} g_{ij}(r_i(u)_{x_i})_{x_j} \right)_{x_i} = 0 \tag{4.1}$$

Stable and convergent LDG methods can de designed for such equations [94]. Let us first look at the simple equation

$$u_t + u_{xxx} = 0.$$

We again rewrite it into a first order system

$$u_t + p_x = 0, \qquad p - q_x = 0, \qquad q - u_x = 0.$$

At this time we follow the idea of LDG methods for convection-diffusion equations and *formally* use the DG method: find $u_h, p_h, q_h \in V_h^k$ such that, for all test functions $v_h, w_h, z_h \in V_h^k$,

$$\int_{I_j} (u_h)_t v_h dx - \int_{I_j} p_h (v_h)_x dx + \hat{p}_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} - \hat{p}_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = 0,$$
$$\int_{I_j} p_h w_h dx + \int_{I_j} q_h (w_h)_x dx - \hat{q}_{j+\frac{1}{2}} (w_h)^-_{j+\frac{1}{2}} + \hat{q}_{j-\frac{1}{2}} (w_h)^+_{j-\frac{1}{2}} = 0,$$
$$\int_{I_j} q_h z_h dx + \int_{I_j} u_h (z_h)_x dx - \hat{u}_{j+\frac{1}{2}} (z_h)^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} (z_h)^+_{j-\frac{1}{2}} = 0.$$

Again, a key ingredient of the design of the LDG method is the choice of the numerical fluxes $\hat{u}$, $\hat{q}$ and $\hat{p}$. Now, the upwinding principle is partially available. After all, the solution with the initial condition $\sin(x)$ is $\sin(x+t)$, hence the wind blows from right to left. The following choice of *alternating* plus *upwinding*

$$\hat{p}_{j+\frac{1}{2}} = p^+_{j+\frac{1}{2}}, \qquad \hat{q}_{j+\frac{1}{2}} = q^+_{j+\frac{1}{2}}, \qquad \hat{u}_{j+\frac{1}{2}} = u^-_{j+\frac{1}{2}},$$

would guarantee stability. The choice is not unique,

$$\hat{p}_{j+\frac{1}{2}} = p^-_{j+\frac{1}{2}}, \qquad \hat{q}_{j+\frac{1}{2}} = q^+_{j+\frac{1}{2}}, \qquad \hat{u}_{j+\frac{1}{2}} = u^+_{j+\frac{1}{2}},$$

would also work. Optimal $(k+1)$-th order L^2 error estimates for not only u but also its derivatives can be proved [93]. Superconvergence to a special projection of the exact solution is proved in [35].

The scheme can be designed for the general nonlinear case along the same lines. For the general multi-dimensional nonlinear case (4.1), we can prove the cell entropy inequality for the square entropy and consequently L^2 stability, just as for the hyperbolic equations [94]. A sub-optimal L^2 error estimate of order $O(h^{k+1/2})$ is also proved in [86].

LDG methods have been designed for the following dispersive wave equations containing higher order (usually odd order) derivatives, usually with stability proof and error estimates:

- PDE with five derivatives [93, 95].
- The $K(m,n)$ equation with *compactons* solutions [48].
- Fifth-order KdV type equations [82].
- Fifth-order fully nonlinear $K(n,n,n)$ equations [82].
- Generalized nonlinear Schrödinger (NLS) equation and the coupled nonlinear Schrödin ger equation [83].
- Kadomtsev-Petviashvili (KP) equation [84].
- Zakharov-Kuznetsov (ZK) equation [84].
- Camassa-Holm (CH) equation [87].
- Hunter-Saxton (HS) equation, its regularization with viscosity and its regularization with dispersion [88, 90].
- Generalized Zakharov system, which is originally introduced to describe the Langmuir turbulence in a plasma [80].
- Degasperis-Procesi (DP) equation [92].

4.2. LDG method for diffusive equations. LDG methods have been designed for the following diffusive equations containing higher even order derivatives, usually with stability proof and error estimates:

- The bi-harmonic type equation and higher even order linear diffusive PDEs [28].
- The Kuramoto-Sivashinsky type equations [85].
- Device simulation models in semi-conductor device simulations: drift-diffusion, hydrodynamic, energy transport, high field, kinetic and Boltzmann-Poisson models [7, 8, 53–55].
- Cahn-Hilliard equation and the Cahn-Hilliard system [78, 79].
- The surface diffusion equation and the Willmore flow [36, 37, 89].

5. Concluding remarks and future work

In this lecture we have given a brief survey for the algorithm formulation, analysis and recent developments and applications of discontinuous Galerkin (DG) methods for solving convection dominated partial differential equations (PDEs). DG methods are very flexible to geometry, boundary condition and h-p adaptivity, and hold a good potential for applications in diverse fields of computational science and engineering. Stable and accurate DG methods can be designed for a wide spectrum of PDEs including conservation laws, convection dominated convection-diffusion equations and dispersive wave equations. Future research is needed for the design of stable DG methods for more nonlinear PDEs in applications, for efficient time discretization (preconditioning, multigrid, exponential type time discretization, deferred correction, etc.), and for a posteriori error estimates to guide adaptivity.

Acknowledgements. The author's research is partially supported by NSF grant DMS-1112700 and DOE grant DE-FG02-08ER25863.

References

[1] Arnold, D. N., *An interior penalty finite element method with discontinuous elements*, SIAM Journal on Numerical Analysis **39** (1982), 742–760.

[2] Bassi, F. and Rebay, S., *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, Journal of Computational Physics **131** (1997), 267–279.

[3] Baumann, C. E. and Oden, J. T., *A discontinuous hp finite element method for convection-diffusion problems*, Computer Methods in Applied Mechanics and Engineering **175** (1999), 311–341.

[4] Biswas, R., Devine K. D., and Flaherty, J., *Parallel, adaptive finite element methods for conservation laws*, Applied Numerical Mathematics **14** (1994), 255–283.

[5] Bramble, J. H. and Schatz, A. H., *High order local accuracy by averaging in the finite element method*, Mathematics of Computation **31** (1977), 94–111.

[6] Burbeau, A., Sagaut P., and Bruneau, Ch. H., *A problem-independent limiter for high-order Runge-Kutta discontinuous Galerkin methods*, Journal of Computational Physics **169** (2001), 111–150.

[7] Cheng, Y., Gamba, I. M., Majorana, A., and Shu, C.-W., *Discontinuous Galerkin solver for Boltzmann-Poisson transients*, Journal of Computational Electronics **7** (2008), 119–123.

[8] _____, *A discontinuous Galerkin solver for Boltzmann Poisson systems in nano devices*, Computer Methods in Applied Mechanics and Engineering **198** (2009), 3130–3150.

[9] Cheng, Y. and Shu, C.-W., *A discontinuous Galerkin finite element method for time dependent partial differential equations with higher order derivatives*, Mathematics of Computation **77** (2008), 699–730.

[10] _____, *Superconvergence and time evolution of discontinuous Galerkin finite element solutions*, Journal of Computational Physics **227** (2008), 9612–9627.

[11] _____, *Superconvergence of local discontinuous Galerkin methods for one-dimensional convection-diffusion equations*, Computers & Structures **87** (2009), 630–641.

[12] _____, *Superconvergence of discontinuous Galerkin and local discontinuous Galerkin schemes for linear hyperbolic and convection diffusion equations in one space dimension*, SIAM Journal on Numerical Analysis **47** (2010), 4044–4072.

[13] Cockburn, B., *Discontinuous Galerkin methods for convection-dominated problems*, in High-Order Methods for Computational Physics, T.J. Barth and H. Deconinck (eds.), Lecture Notes in Computational Science and Engineering, volume 9, Springer, 1999, 69–224.

[14] Cockburn, B. and Guzmán, J., *Error estimates for the Runge-Kutta discontinuous Galerkin method for the transport equation with discontinuous initial data*, SIAM Journal on Numerical Analysis **46** (2008), 1364–1398.

[15] Cockburn, B., Hou, S., and Shu, C.-W., *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case*, Mathematics of Computation **54** (1990), 545–581.

[16] Cockburn, B., Karniadakis, G., and Shu, C.-W., *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, B. Cockburn, G. Karniadakis and C.-W. Shu (eds.), Lecture Notes in Computational Science and Engineering, volume 11, Springer, 2000, Part I: Overview, pp. 3–50.

[17] Cockburn, B., Lin, S.-Y., and Shu, C.-W., *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one-dimensional systems*, Journal of Computational Physics **84** (1989), 90–113.

[18] Cockburn, B., Luskin, M., Shu, C.-W., and Süli, E., *Enhanced accuracy by post-processing for finite element methods for hyperbolic equations*, Mathematics of Computation **72** (2003), 577–606.

[19] Cockburn, B. and Shu, C.-W., *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework*, Mathematics of Computation **52** (1989), 411–435.

[20] ———, *The Runge-Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws*, Mathematical Modelling and Numerical Analysis (M^2AN) **25** (1991), 337–361.

[21] ———, *The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems*, Journal of Computational Physics **141** (1998), 199–224.

[22] ———, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM Journal on Numerical Analysis **35** (1998), 2440–2463.

[23] ———, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, Journal of Scientific Computing **16** (2001), 173–261.

[24] ———, *Foreword for the special issue on discontinuous Galerkin method*, Journal of Scientific Computing, **22-23** (2005), 1–3.

[25] ———, *Foreword for the special issue on discontinuous Galerkin method*, Journal of Scientific Computing **40** (2009), 1–3.

[26] Curtis, S., Kirby, R. M., Ryan, J. K., and Shu, C.-W., *Post-processing for the discontinuous Galerkin method over non-uniform meshes*, SIAM Journal on Scientific Computing **30** (2007), 272–289.

[27] Dawson, C., *Foreword for the special issue on discontinuous Galerkin method*, Computer Methods in Applied Mechanics and Engineering **195** (2006), 3183.

[28] Dong, B. and Shu, C.-W., *Analysis of a local discontinuous Galerkin method for linear time-dependent fourth-order problems*, SIAM Journal on Numerical Analysis **47** (2009), 3240–3268.

[29] Gottlieb, S., Ketcheson, D., and Shu, C.-W., *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*, World Scientific, Singapore, 2011.

[30] Gottlieb, S., Shu, C.-W., and Tadmor, E., *Strong stability-preserving high-order time discretization methods*, SIAM Review **43** (2001), 89–112.

[31] Harten, H., *High resolution schemes for hyperbolic conservation laws*, Journal of Computational Physics **49** (1983), 357–393.

[32] Hesthaven, J. and Warburton, T., *Nodal Discontinuous Galerkin Methods*, Springer, New York, 2008.

[33] Hou, S. and Liu, X.-D., *Solutions of multi-dimensional hyperbolic systems of conservation laws by square entropy condition satisfying discontinuous Galerkin method*, Journal of Scientific Computing **31** (2007), 127–151.

[34] Hu, C. and Shu, C.-W., *Weighted essentially non-oscillatory schemes on triangular meshes*, Journal of Computational Physics **150** (1999), 97–127.

[35] Hufford, C. and Xing, Y., *Superconvergence of the local discontinuous Galerkin method for the linearized Korteweg-de Vries equation*, Journal of Computational and Applied Mathematics **255** (2014), 441–455.

[36] Ji, L. and Xu, Y., *Optimal error estimates of the local discontinuous Galerkin method for Willmore flow of graphs on Cartesian meshes*, International Journal of Numerical Analysis and Modeling **8** (2011), 252–283.

[37] _____, *Optimal error estimates of the local discontinuous Galerkin method for surface diffusion of graphs on Cartesian meshes*, Journal of Scientific Computing **51** (2012), 1–27.

[38] Ji, J., Xu, Y., and Ryan, J., *Negative order norm estimates for nonlinear hyperbolic conservation laws*, Journal of Scientific Computing **54** (2013), 531–548.

[39] _____, *Accuracy-enhancement of discontinuous Galerkin solutions for convection-diffusion equations in multiple-dimensions*, Mathematics of Computation, **81** (2012), 1929–1950.

[40] Jiang, G.-S. and Shu, C.-W., *On cell entropy inequality for discontinuous Galerkin methods*, Mathematics of Computation **62** (1994), 531–538.

[41] _____, *Efficient implementation of weighted ENO schemes*, Journal of Computational Physics **126** (1996), 202–228.

[42] Johnson, J. and Pitkäranta, J., *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Mathematics of Computation **46** (1986), 1–26.

[43] Kanschat, G., *Discontinuous Galerkin Methods for Viscous Flow*, Deutscher Universitätsverlag, Wiesbaden, 2007.

[44] Klockner, A., Warburton, T., Bridge, J., and Hesthaven, J., *Nodal discontinuous Galerkin methods on graphics processors*, Journal of Computational Physics **228** (2010), 7863–7882.

[45] Krivodonova, L., Xin, J., Remacle, J.-F., Chevaugeon, N., and Flaherty, J.E., *Shock detection and limiting with discontinuous Galerkin methods for hyperbolic conservation laws*, Applied Numerical Mathematics **48** (2004), 323–338.

[46] Lesaint, P. and Raviart, P. A., *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor (ed.), Academic Press, 1974, 89–145.

[47] LeVeque, R. J., *Numerical Methods for Conservation Laws*, Birkhauser Verlag, Basel, 1990.

[48] Levy, D., Shu, C.-W., and Yan, J., *Local discontinuous Galerkin methods for nonlinear dispersive equations*, Journal of Computational Physics **196** (2004), 751–772.

[49] Li, B., *Discontinuous Finite Elements in Fluid Dynamics and Heat Transfer*, Birkhauser, Basel, 2006.

[50] Liu, H. and Yan, J., *The direct discontinuous Galerkin (DDG) methods for diffusion problems*, SIAM Journal on Numerical Analysis **47** (2009), 675–698.

[51] _____, *The direct discontinuous Galerkin (DDG) methods for diffusion with interface corrections*, Communications in Computational Physics **8** (2010), 541–564.

[52] Liu, X.-D., Osher, S., and Chan, T., *Weighted essentially non-oscillatory schemes*, Journal of Computational Physics **115** (1994), 200–212.

[53] Liu, Y.-X. and Shu, C.-W., *Local discontinuous Galerkin methods for moment models in device simulations: formulation and one dimensional results*, Journal of Computational Electronics **3** (2004), 263–267.

[54] _____, *Local discontinuous Galerkin methods for moment models in device simulations: Performance assessment and two dimensional results*, Applied Numerical Mathematics **57** (2007), 629–645.

[55] _____, *Error analysis of the semi-discrete local discontinuous Galerkin method for semiconductor device simulation models*, Science China Mathematics **53** (2010), 3255–3278.

[56] Luo, J., Shu, C.-W., and Zhang, Q., *A priori error estimates to smooth solutions of the third order Runge-Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws*, ESAIM: Mathematical Modelling and Numerical Analysis (M^2AN), submitted.

[57] Meng, X., Shu, C.-W., and Wu, B., *Optimal error estimates for discontinuous Galerkin methods based on upwind-biased fluxes for linear hyperbolic equations*, Mathematics of Computation, submitted.

[58] Meng, X., Shu, C.-W., Zhang, Q., and Wu, B., *Superconvergence of discontinuous Galerkin method for scalar nonlinear conservation laws in one space dimension*, SIAM Journal on Numerical Analysis **50** (2012), 2336–2356.

[59] Mirzaee, H., Ji, L., Ryan, J., and Kirby, R. M., *Smoothness-increasing accuracy-conserving (SIAC) post-processing for discontinuous Galerkin solutions over structured triangular meshes*, SIAM Journal on Numerical Analysis **49** (2011), 1899–1920.

[60] Oden, J. T., Babuvska, I., and Baumann, C. E., *A discontinuous hp finite element method for diffusion problems*, Journal of Computational Physics **146** (1998), 491–519.

[61] Qiu, J.-X. and Shu, C.-W., *Hermite WENO schemes and their application as limiters for Runge-Kutta discontinuous Galerkin method: one-dimensional case*, Journal of Computational Physics **193** (2003), 115–135.

[62] _____, *Hermite WENO schemes and their application as limiters for Runge-Kutta discontinuous Galerkin method II: two dimensional case*, Computers & Fluids **34** (2005), 642–663.

[63] ———, *Runge-Kutta discontinuous Galerkin method using WENO limiters*, SIAM Journal on Scientific Computing **26** (2005), 907–929.

[64] ———, *A comparison of troubled-cell indicators for Runge-Kutta discontinuous Galerkin methods using weighted essentially nonoscillatory limiters*, SIAM Journal on Scientific Computing **27** (2005), 995–1013.

[65] Reed, W. H. and Hill, T. R., *Triangular mesh methods for the Neutron transport equation*, Los Alamos Scientific Laboratory Report LA-UR-73-479, Los Alamos, NM, 1973.

[66] Remacle, J.-F., Flaherty, J., and Shephard, M., *An adaptive discontinuous Galerkin technique with an orthogonal basis applied to Rayleigh-Taylor flow instabilities*, SIAM Review **45** (2003), 53–72.

[67] Rivière, B., *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations. Theory and Implementation*, SIAM, Philadelphia, 2008.

[68] Ryan, J. and Shu, C.-W., *On a one-sided post-processing technique for the discontinuous Galerkin methods*, Methods and Applications of Analysis **10** (2003), 295–308.

[69] Ryan, J., Shu, C.-W., and Atkins, H., *Extension of a postprocessing technique for the discontinuous Galerkin method for hyperbolic equations with application to an aeroacoustic problem*, SIAM Journal on Scientific Computing **26** (2005), 821–843.

[70] Shi, J., Hu, C., and Shu, C.-W., *A technique of treating negative weights in WENO schemes*, Journal of Computational Physics **175** (2002), 108–127.

[71] Shu, C.-W., *TVB uniformly high-order schemes for conservation laws*, Mathematics of Computation **49** (1987), 105–121.

[72] ———, *Discontinuous Galerkin methods: general approach and stability*, in Numerical Solutions of Partial Differential Equations, S. Bertoluzza, S. Falletta, G. Russo and C.-W. Shu, Advanced Courses in Mathematics CRM Barcelona, Birkhäuser, Basel, 2009, 149–201.

[73] Shu, C.-W. and Osher, S., *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, Journal of Computational Physics **77** (1988), 439–471.

[74] Smoller, J., *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1994.

[75] Steffan, M., Curtis, S., Kirby, R. M., and Ryan, J., *Investigation of smoothness enhancing accuracy-conserving filters for improving streamline integration through discontinuous fields*, IEEE-TVCG **14** (2008), 680–692.

[76] Wang, C., Zhang, X., Shu, C.-W., and Ning, J., *Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations*, Journal of Computational Physics **231** (2012), 653–665.

[77] Wheeler, M. F., *An elliptic collocation-finite element method with interior penalties*, SIAM Journal on Numerical Analysis **15** (1978), 152–161.

[78] Xia, Y., Xu, Y., and Shu, C.-W., *Local discontinuous Galerkin methods for the Cahn-Hilliard type equations*, Journal of Computational Physics **227** (2007), 472–491.

[79] _____, *Application of the local discontinuous Galerkin method for the Allen-Cahn/Cahn-Hilliard system*, Communications in Computational Physics **5** (2009), 821–835.

[80] _____, *Local discontinuous Galerkin methods for the generalized Zakharov system*, Journal of Computational Physics **229** (2010), 1238–1259.

[81] Xing, X., Zhang, X., and Shu, C.-W., *Positivity preserving high order well balanced discontinuous Galerkin methods for the shallow water equations*, Advances in Water Resources **33** (2010), 1476–1493.

[82] Xu, Y. and Shu, C.-W., *Local discontinuous Galerkin methods for three classes of nonlinear wave equations*, Journal of Computational Mathematics **22** (2004), 250–274.

[83] _____, *Local discontinuous Galerkin methods for nonlinear Schrödinger equations*, Journal of Computational Physics **205** (2005), 72–97.

[84] _____, *Local discontinuous Galerkin methods for two classes of two dimensional nonlinear wave equations*, Physica D **208** (2005), 21–58.

[85] _____, *Local discontinuous Galerkin methods for the Kuramoto-Sivashinsky equations and the Ito-type coupled KdV equations*, Computer Methods in Applied Mechanics and Engineering **195** (2006), 3430–3447.

[86] _____, *Error estimates of the semi-discrete local discontinuous Galerkin method for nonlinear convection-diffusion and KdV equations*, Computer Methods in Applied Mechanics and Engineering **196** (2007), 3805–3822.

[87] _____, *A local discontinuous Galerkin method for the Camassa-Holm equation*, SIAM Journal on Numerical Analysis **46** (2008), 1998–2021.

[88] _____, *Local discontinuous Galerkin method for the Hunter-Saxton equation and its zero-viscosity and zero-dispersion limit*, SIAM Journal on Scientific Computing **31** (2008), 1249–1268.

[89] _____, *Local discontinuous Galerkin method for surface diffusion and Willmore flow of graphs*, Journal of Scientific Computing **40** (2009), 375–390.

[90] _____, *Dissipative numerical methods for the Hunter-Saxton equation*, Journal of Computational Mathematics **28** (2010), 606–620.

[91] _____, *Local discontinuous Galerkin methods for high-order time-dependent partial differential equations*, Communications in Computational Physics **7** (2010), 1–46.

[92] _____, *Local discontinuous Galerkin methods for the Degasperis-Procesi equation*, Communications in Computational Physics **10** (2011), 474–508.

[93] ———, *Optimal error estimates of the semi-discrete local discontinuous Galerkin methods for high order wave equations*, SIAM Journal on Numerical Analysis **50** (2012), 79–104.

[94] Yan, J. and Shu, C.-W., *A local discontinuous Galerkin method for KdV type equations*, SIAM Journal on Numerical Analysis **40** (2002), 769–791.

[95] ———, *Local discontinuous Galerkin methods for partial differential equations with higher order derivatives*, Journal of Scientific Computing **17** (2002), 27–47.

[96] Yang, Y. and Shu, C.-W., *Analysis of optimal superconvergence of discontinuous Galerkin method for linear hyperbolic equations*, SIAM Journal on Numerical Analysis, **50** (2012), 3110-3133.

[97] ———, *Analysis of optimal superconvergence of local discontinuous Galerkin method for one-dimensional linear parabolic equations*, SIAM Journal on Numerical Analysis, submitted.

[98] Zhang, M. and Shu, C.-W., *An analysis of three different formulations of the discontinuous Galerkin method for diffusion equations*, Mathematical Models and Methods in Applied Sciences (M^3AS) **13** (2003), 395–413.

[99] Zhang, Q. and Shu, C.-W., *Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin methods for scalar conservation laws*, SIAM Journal on Numerical Analysis **42** (2004), 641–666.

[100] ———, *Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws*, SIAM Journal on Numerical Analysis **44** (2006), 1703–1720.

[101] ———, *Stability analysis and a priori error estimates to the third order explicit Runge-Kutta discontinuous Galerkin method for scalar conservation laws*, SIAM Journal on Numerical Analysis **48** (2010), 1038–1063.

[102] ———, *Error estimates for the third order explicit Runge-Kutta discontinuous Galerkin method for linear hyperbolic equation in one-dimension with discontinuous initial data*, Numerische Mathematik **126** (2014), 703–740.

[103] Zhang, X. and Shu, C.-W., *On maximum-principle-satisfying high order schemes for scalar conservation laws*, Journal of Computational Physics **229** (2010), 3091-3120.

[104] ———, *On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, Journal of Computational Physics **229** (2010), 8918-8934.

[105] ———, *Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms*, Journal of Computational Physics **230** (2011), 1238-1248.

[106] ———, *Maximum-principle-satisfying and positivity-preserving high order schemes for conservation laws: Survey and new developments*, Proceedings of the Royal Society A **467** (2011), 2752–2776.

[107] Zhang, X., Xia, Y., and Shu, C.-W., *Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes*, Journal of Scientific Computing **50** (2012), 29–62.

[108] Zhang, Y., Zhang, X., and Shu, C.-W., *Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection-diffusion equations on triangular meshes*, Journal of Computational Physics **234** (2013), 295–316.

[109] Zhang, Y.-T. and Shu, C.-W., *Third order WENO scheme on three dimensional tetrahedral meshes*, Communications in Computational Physics **5** (2009), 836–848.

[110] Zhong, X. and Shu, C.-W., *A simple weighted essentially nonoscillatory limiter for Runge-Kutta discontinuous Galerkin methods*, Journal of Computational Physics **232** (2013), 397–415.

[111] Zhu, J., Qiu, J.-X., Shu, C.-W., and Dumbser, M, *Runge-Kutta discontinuous Galerkin method using WENO limiters II: unstructured meshes*, Journal of Computational Physics **227** (2008), 4330–4353.

[112] Zhu, J., Zhong, X., Shu, C.-W., and Qiu, J.-X., *Runge-Kutta discontinuous Galerkin method using a new type of WENO limiters on unstructured meshes*, Journal of Computational Physics **248** (2013), 200–220.

Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
E-mail: shu@dam.brown.edu

Singular stochastic computational models, stochastic analysis, PDE analysis, and numerics

Denis Talay

Abstract. Stochastic computational models are used to simulate complex physical or biological phenomena and to approximate (deterministic) macroscopic physical quantities by means of probabilistic numerical methods. By nature, they often involve singularities and are subject to the curse of dimensionality. Their efficient and accurate simulation is still an open question in many aspects. The aim of this lecture is to review some recent developments concerning the numerical approximation of singular stochastic dynamics, and to illustrate novel issues in stochastic analysis and PDE analysis that they lead to.

Mathematics Subject Classification (2010). Primary 60H30, 60H35,65C05,65C30, 60C35, 65M75; Secondary: 60J55, 60J60.

Keywords. Stochastic numerics; applications of stochastic analysis to partial differential equations and numerical analysis.

1. Introduction

In fields such as biology, ecology, turbulent fluid mechanics, geophysics and environmental sciences, physical laws are not fully known or suffer from the curse of dimensionality. Adding noise to deterministic models or randomizing parameters may not be enough to describe complex phenomena such as cancerous tumor expansions, protein folding, neuron system activity, time evolution of winds, waves, complex flows, movement of groundwater, dynamics of populations, and creation of financial bubbles.

Stochastic computational models (versus models which are fully derived from physical laws) are developed to simulate such phenomena and to approximate (deterministic) macroscopic physical quantities by means of probabilistic numerical methods.

The preceding motivations mean that, by nature, stochastic computational models often inherit singularities from the physical laws they are aimed to mimic. Therefore their efficient and accurate simulation is questionable and actually is still an open question in many aspects.

The numerical analysis of stochastic differential equations with smooth coefficients is now well understood. Optimal convergence rates for efficient numerical methods have been obtained in various theoretical and applied frameworks owing to techniques based on PDE analysis, Malliavin calculus, propagation of chaos and ergodic theories, etc. Although difficult to obtain, these fundamental results in numerical probability are far from being sufficient to tackle singular stochastic computational models.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

The aim of this lecture is to review some recent developments concerning the numerical approximation of singular stochastic dynamics and to illustrate novel issues in stochastic analysis and PDE analysis that they lead to.

2. A short reminder on discretizations of stochastic systems with smooth coefficients

2.1. Standard stochastic differential equations. A stochastic process $(X_t, t > 0)$ is a family of random variables indexed by time. It enjoys the Markov property if

$$\mathbb{E}[g(X_t) \mid X_\theta,\ 0 \le \theta \le s] = \mathbb{E}[g(X_t) \mid X_s]$$

for all bounded measurable functions g and all times $0 \le s \le t$. These Markov processes are key tools to model random physical phenomena and to obtain probabilistic representations of deterministic partial differential equations. Solutions to Brownian stochastic differential equations (SDEs) form a rich class of Markov processes which provide probabilistic interpretations to linear and non-linear parabolic and elliptic PDEs.

Given a vector valued function b and a matrix valued function σ, a weak solution to the Brownian stochastic differential equation with coefficients b and σ is a process (X_t) defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with an increasing family $\{\mathcal{F}_t\}$ of sub-σ-algebras of $\mathcal{F}$ and a Brownian motion (W_t) such that: X_t is 'adapted' i.e. $\mathcal{F}_t$-measurable for all t and, almost surely,

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s,\ \forall t \ge 0. \tag{2.1}$$

The last term in the right-hand side is a stochastic integral. For the construction of stochastic integrals, see, e.g., Revuz and Yor [42]. When $X_0 = x$ a.s., we denote the solution to (2.1) by (X_t^x).

Denote by a the matrix $\sigma \cdot \sigma^*$, where σ^* is the transpose of σ, and by L the second order differential operator

$$L := \sum_k b^k(x)\partial_k + \frac{1}{2}\sum_{j,k} a_k^j(x)\partial_{jk}.$$

Consider the parabolic PDE

$$\frac{d}{dt}u(t,x) = Lu(t,x) \tag{2.2}$$

with initial condition $u(0,x) = f(x)$. Stochastic analysis techniques (stochastic differential calculus, Malliavin calculus, stochastic flows analysis) allow one to prove that the function

$$u(t,x) := \mathbb{E}f(X_t^x) \tag{2.3}$$

is the (classical or viscosity) unique solution to (2.2) and that the flow of the probability distributions μ_t of X_t is a solution in the sense of the distributions to the linear Fokker-Planck equation

$$\frac{d}{dt}\mu_t = L^*\mu_t = -\sum_k \partial_k[b^k(x)\mu_t] + \tfrac{1}{2}\sum_{jk}\partial_{jk}[a_k^j(x)\mu_t], \tag{2.4}$$

notably in cases which are not studied in the classical PDE analysis literature. For example, when the differential operator L is hypoelliptic, Malliavin calculus is a dramatically powerful tool to prove that μ_t has a smooth density for all strictly positive t and to obtain local sharp estimates on partial derivatives of this density (see Kusuoka and Stroock [26]).

These analytical results sustain stochastic numerical methods which combine the approximation of the unknown process (X_t) with an easy to simulate discrete time Markov process $(\overline{X}_t)$, and the approximation of $\mathbb{E}f(\overline{X}_T)$ by means of Monte Carlo methods.

Given a fixed time horizon T, a good candidate for $(\overline{X}_t)$ is the Euler scheme with initial condition X_0 and discretization step $\frac{T}{n}$:

$$\overline{X}_{(p+1)T/n} = \overline{X}_{pT/n} + b(\overline{X}_{pT/n})\tfrac{T}{n} + \sigma(\overline{X}_{pT/n})(W_{(p+1)T/n} - W_{pT/n}),\ p = 0, \ldots, n-1.$$

The simulation of this scheme involves the independent Gaussian random vectors $(W_{(p+1)T/n} - W_{pT/n})$ only. The resulting time discretization error is

$$e_d(n) := \mathbb{E}f(X_T) - \mathbb{E}f(\overline{X}_T).$$

The standard Monte Carlo method consists in approximating

$$\mathbb{E}[f(\overline{X}_T)] \text{ by } \tfrac{1}{N}\sum_{i=1}^{N} f(\overline{X}_T^{(i)}),$$

where the $(\overline{X}_T^{(i)})$'s are independent samples of $\overline{X}_T$. The resulting statistical error is

$$e_s(n, N) := \mathbb{E}f(\overline{X}_T) - \frac{1}{N}\sum_{i=1}^{N} f(\overline{X}_T^{(i)}).$$

The statistical error $e_s(n, N)$ can be estimated by using non-asymptotic versions of the central limit theorem (e.g., the Bikelis theorem [23, 40]): for example, under mild assumptions one has

$$\exists C > 0,\ \forall n > T,\ \forall N \geq 1,\ \ \mathbb{E}|e_s(n, N)| \leq \frac{C}{\sqrt{N}}. \tag{2.5}$$

For more precise estimates, see Section 7.

In various contexts ([3, 21, 27, 37, 44]) the discretization error $e_d(n)$ can be expanded w.r.t. n:

$$e_d(n) = \frac{C_1}{n^{K_1}} + \frac{C_2}{n^{K_2}} + \frac{C_3}{n^{K_3}} + \ldots + \mathcal{O}\left(\frac{1}{n^{K_m}}\right). \tag{2.6}$$

This justifies the use of low numerical cost Romberg-Richardson extrapolation procedures to exponentially decrease the time discretization error: see [44].

The following formal calculation gives an intuition for the equality (2.3) and for the methodology to get (2.6). To simplify, suppose that b, σ and f are bounded and of class $\mathcal{C}^\infty$ with bounded derivatives. Then $u(t, x)$ enjoys the same properties.

Notice first that, as the Brownian motion has independent Gaussian increments,

$$\mathbb{E}[\overline{X}_{(p+1)T/n} - \overline{X}_{pT/n}] = \mathbb{E}b(\overline{X}_{pT/n})\ \tfrac{T}{n},$$
$$\mathbb{E}[(\overline{X}_{(p+1)T/n} - \overline{X}_{pT/n}) \cdot (\overline{X}_{(p+1)T/n} - \overline{X}_{pT/n})^*] = \mathbb{E}a(\overline{X}_{pT/n})\ \tfrac{T}{n} + \mathcal{O}\left(\tfrac{1}{n^2}\right).$$

Denoting by $\overline{X}^x_{pT/n}$ the Euler scheme with initial condition $\overline{X}_0 = x$, we thus have

$$\begin{aligned}
&\mathbb{E}f(\overline{X}^x_T) - u(T,x)\\
&\quad= \sum_{p=0}^{n-1} \mathbb{E}\left[u(T-(p+1)\tfrac{T}{n}, \overline{X}^x_{(p+1)T/n}) - u(T-p\tfrac{T}{n}, \overline{X}^x_{pT/n})\right]\\
&\quad= \sum_{p=0}^{n-1} \mathbb{E}\left[u(T-(p+1)\tfrac{T}{n}, \overline{X}^x_{pT/n}) - u(T-p\tfrac{T}{n}, \overline{X}^x_{pT/n})\right]\\
&\qquad+ \frac{T}{n}\sum_{p=0}^{n-1} \mathbb{E}Lu(T-(p+1)\tfrac{T}{n}, \overline{X}^x_{pT/n}) + \sum_{p=0}^{n-1} \mathcal{O}\left(\tfrac{1}{n^2}\right)\\
&\quad= \frac{T}{n}\sum_{p=0}^{n-1} \mathbb{E}\left[-\frac{\partial u}{\partial t}(T-p\tfrac{T}{n}, \overline{X}^x_{pT/n}) + Lu(T-p\tfrac{T}{n}, \overline{X}^x_{pT/n})\right] + \sum_{p=0}^{n-1} \mathcal{O}\left(\tfrac{1}{n^2}\right)\\
&\quad= \mathcal{O}\left(\tfrac{1}{n}\right),
\end{aligned} \tag{2.7}$$

since $\frac{\partial u}{\partial t}(t,x) = Lu(t,x)$. In the preceding equalities the main difficulty is hidden: one has to justify that the remaining terms are of the prescribed order w.r.t. n. This requires accurate pointwise estimates on the partial derivatives of $u(t,x)$.

2.2. McKean-Vlasov stochastic differential equations. Stochastic particle systems with McKean–Vlasov interactions arise in physics, fluid mechanics, economy, biology, etc.

Given N independent Brownian motions $(W^{(i)}_t)$, multi-dimensional coefficients B and S, and McKean interaction kernels b and σ, consider the following system

$$\begin{aligned}
X^{(i)}_t = X^{(i)}_0 &+ \int_0^t B(s, X^{(i)}_s, \textstyle\int b(X^{(i)}_s, y)\overline{\nu}^N_s(dy))ds\\
&+ \int_0^t S(s, X^{(i)}_s, \textstyle\int \sigma(X^{(i)}_s, y)\overline{\nu}^N_s(dy))dW^{(i)}_s,
\end{aligned} \tag{2.8}$$

where $\overline{\nu}^N_s$ is the marginal distribution at time s of the empirical distribution $\overline{\nu}^N$ of the trajectories of the particles

$$\bar{\nu}^N := \frac{1}{N}\sum_{j=1}^N \delta_{X^{(j)}}.$$

Notice that the processes $(X^{(i)}_t)$ are dependent. However, seminal works by McKean and Sznitman [43] show that the particle system propagates chaos in the sense that the probability distribution of $\bar{\nu}^N$ converges weakly when N goes to infinity. The limit distribution is concentrated at the probability law of the process (X_t), solution to the McKean-Vlasov SDE

$$\begin{cases}
X_t = X_0 + \int_0^t B(s, X_s, \int b(X_s, y)\nu_s(dy))ds\\
\qquad\quad + \int_0^t S(s, X_s, \int \sigma(X_s, y)\nu_s(dy))dW_s,\\
\nu_s(dy) := \text{probability distribution of } X_s.
\end{cases} \tag{2.9}$$

In addition, the flow of the probability distributions ν_t solves the non-linear McKean-Vlasov-Fokker-Planck equation

$$\frac{d}{dt}\nu_t = L^*_{\nu_t}\nu_t, \tag{2.10}$$

where, A denoting the matrix $S \cdot S^*$, L_ν^* is the formal adjoint of the non-linear differential operator

$$L_\nu := \sum_k B^k(t,x,\int b(x,y)\nu(dy))\partial_k + \frac{1}{2}\sum_{j,k} A_k^j(t,x,\int \sigma(x,y)\nu(dy))\partial_{jk}. \qquad (2.11)$$

This construction has important analytical and numerical consequences.

From an analytical point of view, the McKean-Vlasov SDEs (2.9) allow one to construct probabilistic interpretations for a wide family of macroscopic equations including smoothed versions of the Navier-Stokes and Boltzmann equations. The theory is well developed for smooth functions B, S, and smooth McKean interaction kernels b, σ, and also for some particular irregular kernels, often under strong ellipticity conditions on the differential operator L_ν (see for example Sznitman's survey [43], Osada [39], Méléard [36]).

From a numerical point point of view, whereas the time discretization of (X_t) does not lead to an algorithm since ν_t is unknown, the particle system $\{(X_t^{(i)}), i = 1,\dots,N\}$ is a Markov process which can be discretized in time (e.g., by using the Euler scheme) and thus simulated: for all T, the solution ν_T to (2.10) is approximated by the empirical distribution of the simulated particles at time T.

When the functions B, S, b, σ are smooth, optimal convergence rates have been obtained, e.g. in [2, 6, 10]. For example, given a differentiable function Π, consider the scalar conservation law

$$\frac{\partial V}{\partial t}(t,x) = \frac{1}{2}\frac{\partial^2 V}{\partial x^2}(t,x) - \frac{\partial}{\partial x}\Pi \circ V(t,x).$$

A formal identification of $V(t,x)$ as the distribution function of the solution ν_t to (2.10) leads to the following particle system:

$$X_t^{(i)} = X_0^{(i)} + \int_0^t \Pi'\left(\frac{1}{N}\sum_{j=1}^N H(X_s^{(i)} - X_s^{(j)})\right) ds + W_t^{(i)}, \qquad (2.12)$$

where H is the Heaviside function. Let $\overline{X}_{pT/n}^{(i)}$ be the Euler discretization of the system (2.12). Set

$$\overline{V}(T,x) := \frac{1}{N}\sum_{i=1}^N H(x - \overline{X}_T^{(i)}).$$

In [7] the following error estimate is obtained:

Theorem 2.1. *Suppose that the $X_0^{(i)}$ are independent and have the same twice continuously differentiable probability distribution function $V_0(x)$ satisfying $V_0'(x) \le Ke^{-cx^2}$ for some strictly positive constants c and K.*

Suppose also that Π is of class $\mathcal{C}^3(\mathbb{R})$. Then

$$\exists C > 0,\ \forall n > T,\ \forall N \ge 1,$$

$$\sup_{x\in\mathbb{R}} |V(T,x) - \overline{V}(T,x)| + |V(T,\cdot) - \overline{V}(T,\cdot)|_{L^1(\mathbb{R})} \le \frac{C}{n} + \frac{C}{\sqrt{N}}.$$

Notice that the statistical error is of order $\frac{1}{\sqrt{N}}$ as in the case when the particles are independent (cf. (2.5)) and that the discretization error is of order $\frac{1}{n}$ as in the case of standard SDEs (cf. (2.7)).

3. SDEs with discontinuous coefficients

In [28] Equation (2.1) is considered with bounded measurable drift coefficient b and continuous diffusion coefficient σ. The only additional assumption is that the operator L is uniformly strongly elliptic, that is,

$$\exists 0 < \lambda < \Lambda,\ \lambda|\xi|^2 \leq \xi^* a(x)\xi \leq \Lambda|\xi|^2,\ \forall \xi. \tag{3.1}$$

No convergence rate analysis from the literature can be applied in this framework because of the lack of regularity of the coefficients: estimates for partial derivatives of $u(t,x)$ cannot be obtained by classical PDE analysis techniques, Malliavin calculus, or differentiation of stochastic flows.

Instead, the technique in [28] consists at smoothing the drift coefficient and discretizing the smoothed SDE. Let (X_t^ϵ) be the solution to

$$X_t^\epsilon = X_0^\epsilon + \int_0^t b_\epsilon(X_s^\epsilon)ds + \int_0^t \sigma(X_s^\epsilon)dW_s,\ \forall t \geq 0.$$

Suppose that, given two families of functions $\mathbb{F}$ and $\mathbb{M}$, for all test functions f in $\mathbb{F}$ and smoothed coefficients B^ϵ in $\mathbb{M}$ one has

$$|\mathbb{E}f(X_T) - \mathbb{E}f(X_T^\epsilon)| \leq C\epsilon^\gamma$$

and

$$|\mathbb{E}f(X_T^\epsilon) - \mathbb{E}f(\overline{X}_T^\epsilon)| \leq \frac{C}{n^\delta \epsilon^\beta}$$

for some constants C, γ, β and δ depending on T, $\mathbb{F}$ and $\mathbb{M}$ uniquely. Then, for some possibly new positive number C,

$$|\mathbb{E}f(X_T) - \mathbb{E}f(\overline{X}_T^\epsilon)| \leq \frac{C}{n^\kappa}$$

with $\kappa = \delta - \frac{\delta\beta}{\gamma+\beta}$.

The authors exhibit several classes of functions $\mathbb{F}$ and $\mathbb{M}$ for which the above conditions hold true. In short, suitable functions b_ϵ approximate B in L^p norm for some $p > 1$; suitable functions f are those which satisfy

$$\exists c > 0,\ \lim_{|x|\to\infty} |f(x)|e^{-c|x|^2} = 0$$

and, for some r large enough,

$$\mathbb{E}\int_0^T |\nabla u(s, X_s)|^r ds < \infty,$$

where $u(t,x)$ is the solution to the PDE (2.2).

In addition, if b_ϵ, a and f are of class $\mathcal{C}^3(\mathbb{R})$ then

$$|\mathbb{E}f(X_T^\epsilon) - \mathbb{E}f(\overline{X}_T^\epsilon)| \leq \frac{C}{n},$$

where C depends on the L^∞ norms of b_ϵ and its partial derivatives up to order 3.

Another interesting approach is due to Alfonsi [1] for the particular case of the Cox-Ingersoll-Ross (CIR) model in financial mathematics. See also Deaconu and Herrmann [16] for the construction and analysis of an extension of the Walk on Spheres method to approximate hitting times of the CIR process.

Other techniques may be used to estimate the effects of smoothing the coefficients around singularities. We here present a useful result to localize the discontinuities of the coefficients. More general results hold true: see Bossy et al. [4].

Theorem 3.1. *Suppose that the functions b and σ are bounded.*

Let g be a positive and increasing function in $\mathcal{C}^1([0,T);\mathbb{R}^+)$ such that g^α is integrable on $[0,T)$ for all $1 \leq \alpha < 2$. Suppose also that there exists $1 < \beta < 1+\eta$, where $\eta := \frac{1}{4(|B|_\infty \wedge 1)^4}$, such that

$$\int_0^T g^{2\beta-1}(v) g'(v) \frac{(T-v)^{1+\eta}}{v^\eta}\, dv < +\infty.$$

Then there exists a constant C, depending only on β, K and T, such that, for all vector ξ, real number $0 < \varepsilon < 1/2$, and integer n large enough,

$$\frac{1}{n}\sum_{p=0}^{n} \mathbb{P}\left[|X_{pT/n} - \xi| \leq \frac{1}{n^{1/2-\varepsilon}}\right] g(pT/n) \leq \frac{C}{n^{1/2-\varepsilon}}$$

and

$$\frac{1}{n}\sum_{p=0}^{n} \mathbb{P}\left[|\overline{X}_{pT/n} - \xi| \leq \frac{1}{n^{1/2-\varepsilon}}\right] g(pT/n) \leq \frac{C}{n^{1/2-\varepsilon}}.$$

As noticed in [34] the constraint on n is that

$$\exp(-n^\varepsilon) \leq \frac{C}{n^{3/2-\varepsilon}}.$$

An example of a suitable function g is $g(t) = \frac{1}{\sqrt{T-t}}$. This is of interest since (remember the comment after (2.7)) typically one would like to take g as a suitable norm of a partial derivative of $u(t,x)$ which has this type of singularity in time when the data of the PDE (2.2) are irregular.

4. SDEs with weighted local times and interface PDE

Many physical conservation laws involve operators of the type $\nabla.(a(x)\nabla v(x))$ where $a(x)$ is a discontinuous function along hypersurfaces: transport equations in geophysics, Poisson-Boltzmann equations in molecular dynamics, diffraction problems, etc.

From a stochastic point of view the situation differs from the preceding section since a formal expansion of the definition of the operator leads to the definition of the coefficient B as a singular measure rather than a function. Dirichlet form theory, Itô–Fukushima's decomposition and Portenko's approach involve abstract processes whose numerical simulation does not seem possible: see Lejay's survey [30].

We thus here follow another approach for which [30] is a good introduction. To this end, we need to introduce the notion of local time. For all process $Z := (Z_t)$ which can be written as

$$Z_t = Z_0 + \int_0^t \phi_s dW_s + A_t^+ - A_t^-$$

for some adapted process (ϕ_t) and some adapted continuous increasing processes (A_t^+) and (A_t^-), the right-sided local time $L_t^\xi(Z)$ of Z at point ξ is the increasing continuous process such that

$$|Z_t - \xi| = |Z_0 - \xi| + \int_0^t \mathrm{sgn}(Z_s - \xi) dZ_s + L_t^\xi(Z),$$

where $\mathrm{sgn}(x) := 1$ for $x > 0$ and $\mathrm{sgn}(x) := -1$ for $x \leq 0$ (see, e.g., Revuz and Yor [42]). At fixed ξ, the Stieljes measure in t, $dL_t^\xi(Z)$, is carried by the set $\{t;\ Z_t = \xi\}$:

$$\int_0^\infty \mathbb{I}_{Z_s \neq \xi} dL_s^\xi(Z) = 0.$$

In addition, almost surely

$$L_t^\xi(Z) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_0^t \phi_s^2\, \mathbb{I}_{\xi \leq X_s \leq \xi + \epsilon} ds. \tag{4.1}$$

4.1. The one-dimensional case. The results of this section come from Martinez and Talay [34].

Consider the real valued function $a(x) = (\sigma(x))^2$ defined on $\mathbb{R}$ and the interface (or diffraction) PDE with boundary transmission condition

$$\begin{cases} \partial_t u(t,x) - \frac{1}{2}\partial_x(a(x)\partial_x u(t,x)) = 0,\ (t,x) \in (0,T] \times (\mathbb{R} - \{0\}), \\ u(t,0+) = u(t,0-),\ t \in [0,T], \\ u(0,x) = f(x),\ x \in \mathbb{R}, \\ a(0+)\partial_x u(t,0+) = a(0-)\partial_x u(t,0-),\ t \in [0,T]. \end{cases} \tag{4.2}$$

Assume the uniform strong ellipticity condition

$$\exists 0 < \lambda < \Lambda,\ \lambda \leq a(x) = (\sigma(x))^2 \leq \Lambda,\ \forall\, x \in \mathbb{R}. \tag{4.3}$$

Consider the one-dimensional SDE with weighted local time

$$X_t = x + \int_0^t \sigma(X_s) dW_s + \int_0^t \sigma(X_s)\sigma'_-(X_s) ds + \frac{a(0+) - a(0-)}{2a(0+)} L_t^0(X), \tag{4.4}$$

where σ'_- is the left derivative of σ.

By considering the SDE (4.4), one can prove the existence and uniqueness of smooth solutons to (4.2):

Theorem 4.1. *Let us assume condition (4.3) and that the function σ is of class $\mathcal{C}_b^3(\mathbb{R} - \{0\})$. Moreover, we assume that σ and its derivatives have finite left and right limits at 0. Let (X_t) be the solution to (4.4). Let the bounded function f be in the set*

$$\begin{aligned} \mathcal{W}^2 := \{ & g \in \mathcal{C}_b^2(\mathbb{R} - \{0\}),\ g^{(i)} \in L^2(\mathbb{R}) \cap L^1(\mathbb{R}) \text{ for } i = 1,2; \\ & a(0+)g'(0+) = a(0-)g'(0-)\}. \end{aligned} \tag{4.5}$$

Then the function

$$u(t,x) := \mathbb{E}f(X_t^x),\ (t,x) \in [0,T] \times \mathbb{R},$$

is the unique solution in $\mathcal{C}_b^{1,2}([0,T] \times (\mathbb{R} - \{0\})) \bigcap \mathcal{C}^0([0,T] \times \mathbb{R})$ *to (4.2).*

The next pointwise estimates on the partial derivatives of $u(t,x)$ are crucial to analyze convergence rates for time discretizations of (4.4).

Theorem 4.2. *In addition to the hypotheses made in Theorem 4.1, suppose that the function* σ *is of class* $\mathcal{C}_b^4(\mathbb{R} - \{0\})$ *and that its three first derivatives have finite left and right limits at* 0*. Set*

$$\begin{aligned}\mathcal{W}^4 :=& \{g \in \mathcal{C}_b^4(\mathbb{R} - \{0\}),\ g^{(i)} \in L^2(\mathbb{R}) \cap L^1(\mathbb{R}) \text{ for } i = 1,\dots,4;\\ & a(0+)g'(0+) = a(0-)g'(0-) \text{ and } a(0+)(\mathcal{L}g)'(0+) = a(0-)(\mathcal{L}g)'(0-)\},\end{aligned} \tag{4.6}$$

where, for all $x \neq 0$*,*

$$\mathcal{L}g(x) := \sigma(x)\sigma'(x)\partial_x g(x) + \frac{1}{2}a(x)\partial_{xx}^2 g(x). \tag{4.7}$$

Then, for all $j = 0, 1, 2$ *and* $i = 1,\dots,4$ *such that* $2j + i \leq 4$*,*

$$\exists C > 0,\ \forall x \in \mathbb{R},\ \forall t \in (0,T],\ \forall f \in \mathcal{W}^4,\ |\partial_t^j \partial_x^i u(t,x)| \leq \frac{C}{\sqrt{t}}, \tag{4.8}$$

where the constant C *only depends on* T *and the* $L^1(\mathbb{R})$ *norm of* $f^{(i)}$ $(1 \leq i \leq 3)$*.*

A transformed Euler scheme. The numerical approximation of the process $L_t^0(X)$ is a critical issue: on the one hand, it is the local time of the unknown process (X_t); on the other hand, Equality (4.1) shows that time discretizations of local times are numerically unstable.

We thus apply a transformation introduced by Le Gall [29] to get existence, uniqueness and the Markov property for the solution to an equation more general than (4.4). This one-to-one transformation leads to a new stochastic differential equation with discontinuous coefficients but without local time, which can be discretized by the standard Euler scheme.

In our context, set

$$\beta_+ := \tfrac{2a(0-)}{a(0+)+a(0-)} \text{ and } \beta_- := \tfrac{2a(0+)}{a(0+)+a(0-)}, \tag{4.9}$$

and

$$\begin{cases}\beta(x) := x(\beta_- \mathbb{I}_{x\leq 0} + \beta_+ \mathbb{I}_{x>0}),\\ \beta^{-1}(x) := \frac{x}{\beta_-}\mathbb{I}_{x\leq 0} + \frac{x}{\beta_+}\mathbb{I}_{x>0}.\end{cases} \tag{4.10}$$

Set also

$$\begin{cases}\tilde{\sigma}(x) &:= \sigma \circ \beta^{-1}(x)(\beta_- \mathbb{I}_{x\leq 0} + \beta_+ \mathbb{I}_{x>0}),\\ \tilde{b}(x) &:= \sigma \circ \beta^{-1}(x)\sigma'_- \circ \beta^{-1}(x)(\beta_- \mathbb{I}_{x\leq 0} + \beta_+ \mathbb{I}_{x>0}).\end{cases} \tag{4.11}$$

From Itô–Tanaka's formula (see, e.g., Revuz and Yor [42, Chap.VI]) applied to $\beta(X_t)$ we see that the process $Y := \beta(X)$ satisfies the SDE with discontinuous coefficients:

$$Y_t = \beta(X_0) + \int_0^t \tilde{\sigma}(Y_s)dB_s + \int_0^t \tilde{b}(Y_s)ds. \tag{4.12}$$

Let $\overline{Y}$ be the Euler approximation of (Y_t), and the transformed Euler scheme for (X_t) be defined as

$$\overline{X}_{pT/n} = \beta^{-1}(\overline{Y}_{pT/n}). \tag{4.13}$$

We have the following convergence rate result.

Theorem 4.3. *Under the hypotheses made in Theorem 4.2, there exists a positive number C such that, for all initial conditions f in $\mathcal{W}^4$, all $0 < \epsilon < \frac{1}{2}$ and all n large enough,*

$$|\mathbb{E}f(X_T) - \mathbb{E}f(\overline{X}_T)| \leq \frac{C}{n^{(1-\epsilon)/2}}. \tag{4.14}$$

Random walk methods. Other numerical methods have recently been developed which involve a space discretization, random walks on the grid, and flips of a coin at point 0 to mimic the effect of the weighted local time in (4.4). For their convergence rate analysis, see, e.g., Etoré [20] and Lejay and Martinez [31].

4.2. The linear 3D Poisson-Boltzmann PDE in molecular dynamics. The results in this section come from Bossy et al. [11].

The Poisson-Boltzmann PDE in molecular dynamics describes the electrostatic potential around a biomolecular assembly and is used to compute the solvatation free energy and the electrostatic forces exerted by the solvent on the molecule. In its linearized version, it reads

$$\begin{cases} -\nabla \cdot (\varepsilon(x)\nabla u(x)) + \kappa^2(x)u(x) = \sum_{i=1}^N q_i \delta_{x_i}, & x \in \mathbb{R}^3, \\ \varepsilon_{\text{int}} \nabla^{\text{int}} u(y) \cdot n(y) = \varepsilon_{\text{ext}} \nabla^{\text{ext}} u(y) \cdot n(y), \ y \in \Gamma, \end{cases} \tag{4.15}$$

where $\varepsilon(x)$ is the permittivity of the medium, $\kappa^2(x)$ is the ion accessibility parameter, and $x_1, \ldots, x_N$ are the positions of the atoms in the molecule with charges q_i. We here deal with the simplified coefficients and geometry

$$\varepsilon(x) := \begin{cases} \varepsilon_{\text{int}} > 0 & \text{if } x \in \overline{\Omega_{\text{int}}}, \\ \varepsilon_{\text{ext}} > 0 & \text{if } x \in \Omega_{\text{ext}}, \end{cases} \qquad \kappa(x) = \begin{cases} 0 & \text{if } x \in \overline{\Omega_{\text{int}}}, \\ \bar{\kappa} > 0 & \text{if } x \in \Omega_{\text{ext}}, \end{cases}$$

Ω_{int} and Ω_{ext} being two open subsets of $\mathbb{R}^3$. We suppose that Ω_{int} is bounded with boundary Γ, $\Omega_{\text{int}} \cap \Omega_{\text{ext}} = \emptyset$, and $\overline{\Omega}_{\text{int}} \cup \overline{\Omega}_{\text{ext}} = \mathbb{R}^3$. To formulate the boundary condition we have denoted by $n(y)$ the unit outward normal to Γ at y in Γ, and set

$$\nabla^{\text{int}}\varphi(x) := \lim_{y \in \Omega_{\text{int}},\, y \to x} \nabla\varphi(y) \quad \text{and} \quad \nabla^{\text{ext}}\varphi(x) := \lim_{y \in \Omega_{\text{ext}},\, y \to x} \nabla\varphi(y), \quad \forall x \in \Gamma.$$

Let χ be a $\mathcal{C}^\infty$ function with compact support in Ω_{int} such that $\chi(x) = 1$ in the neighborhood of the points $\{x_1, \ldots, x_N\}$. Consider the function

$$G(x) := \sum_i \frac{1}{4\pi} \frac{q_i}{\epsilon_{\text{int}}} \frac{1}{|x - x_i|}, \qquad x \in \mathbb{R}^3.$$

The function $v := u - \chi G$ solves the Poisson-Boltzman equation with regularized source term

$$-\nabla \cdot (\epsilon(x)\nabla v(x)) + \kappa^2(x)v(x) = g(x), \qquad x \in \mathbb{R}^3, \tag{4.16}$$

where here

$$g(x) = \epsilon_{\text{int}} \left(G(x)\Delta\chi(x) + \nabla G(x) \cdot \nabla\chi(x) \right).$$

Assume that Γ is a smooth manifold of class $\mathcal{C}^3$. Denote by $\pi(x)$ the orthogonal projection of x on Γ and by $\rho(x)$ the signed distance between x and Γ, that is, $\rho(x) := (x - \pi(x)) \cdot n(\pi(x))$.

The following theorem is the foundation of the probabilistic interpretation of the linear and non-linear Poisson–Boltzmann equations. The technical difficulties of its proof come from the fact that the dynamics of the unknown process (X_t) depends on the local time of the auxiliary process $(\rho(X_t))$.

Theorem 4.4. *The SDE with weighted local time*

$$\begin{cases} X_t &= x + \int_0^t \sqrt{2\epsilon(X_s)} dW_s + \dfrac{\epsilon_{\text{ext}} - \epsilon_{\text{int}}}{2\epsilon_{\text{ext}}} \int_0^t n(X_s) dL_s^0(Y), \\ Y_t &= \rho(X_t), \end{cases} \tag{4.17}$$

where $L_t^0(Y)$ is the right-sided local time at 0 of the process (Y_t), has a unique weak solution.

One then can prove the following result which extends Theorem 4.1 to Poisson-Boltzmann equation.

Theorem 4.5. *Let g be a smooth function and v be the solution to (4.16). Then, for all $x \in \mathbb{R}^3$,*

$$v(x) = \mathbb{E}\left[\int_0^{+\infty} g(X_t^x) \exp\left(- \int_0^t \kappa^2(X_s^x) ds \right) dt \right]. \tag{4.18}$$

The key ingredient for the preceding theorem is the following, which extends the classical Itô-Meyer formula.

Proposition 4.6. *For all functions φ in $\mathcal{C}_b^0(\mathbb{R}^d) \cap \mathcal{C}_b^2(\mathbb{R} \setminus \Gamma)$ such that*

$$\varepsilon_{\text{int}} \nabla^{\text{int}} \varphi(x) \cdot n(x) = \varepsilon_{\text{ext}} \nabla^{\text{ext}} \varphi(x) \cdot n(x), \quad \forall x \in \Gamma,$$

one has

$$\varphi(X_t) = \varphi(X_0) + \int_0^t \mathbb{I}_{X_s \notin \Gamma} \sqrt{2\varepsilon(X_s)} \nabla\varphi(X_s) \cdot dB_s + \int_0^t \mathbb{I}_{\{X_s \notin \Gamma\}} \mathcal{L}\varphi(X_s) ds,$$

where $\mathcal{L}\varphi(x) := \nabla \cdot (\varepsilon(x) \nabla\varphi(x))$.

The stochastic representation (4.18) does not suffice for the construction of an efficient stochastic numerical method to solve the Poisson–Boltzmann equation because the approximation of $L_t^0(Y)$ and thus the discretization of (X_t) is a critical issue. However the process (X_t) allows us to exhibit another representation which is open to the derivation of numerical methods.

For $h > 0$ define the following sequence of random times

$$\begin{aligned} \tau_k &= \inf\{t \geq \tau'_{k-1} : \rho(X_t^x) = -h\}, \\ \tau'_k &= \inf\{t \geq \tau_k : X_t^x \in \Gamma\}. \end{aligned}$$

Since $\Delta(u-G)=0$ in Ω_{int}, for all x such that $\rho(x)\le -h$,

$$u(x)=\mathbb{E}[u(X^x_{\tau'_1})-G(X^x_{\tau'_1})]+G(x).$$

For all $x\in\Omega_{\text{ext}}$,

$$u(x)=\mathbb{E}\left[u(X^x_{\tau_1})\exp\left(-\int_0^{\tau_1}\kappa^2(X^x_t)dt\right)\right].$$

Recursively applying the two preceding formulas leads to the following result.

Theorem 4.7. *One has*

$$u(x)=\mathbb{E}\left[\sum_{k=1}^{+\infty}(G(X^x_{\tau_k})-G(X^x_{\tau'_k}))\exp\Big(-\int_0^{\tau_k}\kappa^2(X^x_t)dt\Big)\right].$$

When $\kappa(x)$ and $a(x)$ are constant in Ω_{int} and Ω_{ext} (which implies that (X_t) behaves as a Brownian motion outside neighborhoods of Γ), the preceding formula justifies the Walk on Spheres algorithm introduced in this context by Mascagni and Simonov [35], which is based on the sampling of (τ_k, W_{τ_k}). It also allows one to get accurate convergence rate estimates in terms of h: see [11].

4.3. The general multi-dimensional case. Consider the differential operator

$$\mathcal{L}v(x):=\frac{1}{2}\nabla\cdot(a(x)\nabla v(x))+b(x)\nabla v(x), \tag{4.19}$$

and the general interface problem with transmission boundary condition at the boundary Γ of a bounded domain D in $\mathbb{R}^d$:

$$\begin{cases}\partial_t u(t,x)-\mathcal{L}u(t,x)=0,\ (t,x)\in(0,T]\times(\mathbb{R}^d\setminus\Gamma),\\ u(0,x)=f(x),\ x\in\mathbb{R}^d,\\ [u(t,x)]=0,\ (t,x)\in(0,T]\times\Gamma,\\ [n(\pi(x))^t a(x)\nabla u(t,x)]=0,\ (t,x)\in(0,T]\times\Gamma,\end{cases} \tag{4.20}$$

where $[f(x)]:=\lim_{y\to x,y\in\Omega_{\text{ext}}}f(y)-\lim_{y\to x,y\in\Omega_{\text{int}}}f(y)$.

Suppose that $a(x)$ is continuous except along Γ with finite limits $a^{\text{ext}}(x)$ and $a^{\text{int}}(x)$ on each side of Γ. Consider the SDE with weighted local time

$$\begin{cases}X_t &= x+\int_0^t b(X_s)ds+\int_0^t\sigma(X_s)dW_s\\ &\quad+\int_0^t\dfrac{(a^{\text{ext}}(\pi(X_s))-a^{\text{int}}(\pi(X_s)))n(\pi(X_s))}{n(\pi(X_s))^t(a^{\text{ext}}(\pi(X_s))+a^{\text{int}}(\pi(X_s)))n(\pi(X_s))}dL^0_s(Y),\\ Y_t &:=\rho(X_t),\end{cases} \tag{4.21}$$

where ρ is the signed distance to the surface Γ. The next theorem comes from Niklitschek-Soto and Talay [38].

Theorem 4.8. *Let $\Gamma\subset\mathbb{R}^d$ be a bounded simply connected manifold of class $\mathcal{C}^3$. Suppose that the $\sigma^i_j(x)$ and $b^i(x)$ respectively are of class $\mathcal{C}^3_b(\mathbb{R}^d-\Gamma)$ and $\mathcal{C}^2_b(\mathbb{R}^d-\Gamma)$, these functions and their partial derivatives having finite limits on each side of Γ. Suppose also that $a(x)$ satisfies the strong ellipticity condition (3.1). Then there exists a unique weak solution to the SDE (4.21) and the function $u(t,x):=\mathbb{E}f(X^x_t)$ is the unique solution to (4.20) in a space which is the multi-dimensional version of (4.6).*

The construction of a numerically efficient discretization of (4.21) is in progress: the simple one-to-one transformation β in Section 4.1 has no multi-dimensional equivalent and the Walk on Spheres numerical method does not extend to non locally constant functions $a(x)$ and $\kappa(x)$.

5. Stochastic computational models for complex flows in boundary layers

Many stochastic computational models have been designed to take into account the fact that the Reynolds number for flows is close to 0 in the vicinity of boundaries. We here focus on some particular models.

5.1. Stochastic Lagrangian models. In the statistical approach of turbulent flows, the velocity $U(t,x)$, the pressure, and other fundamental quantities, are random fields which are described by their Reynolds decomposition: for example, the Reynolds decomposition of the velocity field writes

$$U(t,x,\omega) = \langle U\rangle(t,x) + \mathbf{u}(t,x,\omega),$$

where the Reynolds average part $\langle U\rangle$ is deterministic, and $\mathbf{u}$ is the fluctuating part. To compute the average and higher moments of the velocity field, one needs to model the average part and moments of the fluctuating part. Pope's approach to this modelling issue consists at describing, through a stochastic model, the Lagrangian properties of the flow. In a series of papers initiated in the eighties, S. Pope has proposed Lagrangian stochastic models to describe the position X_t and the instantaneous velocity U_t of a fluid particle. Depending on the flow, other Lagrangian characteristics of the turbulence are added to the model. For a fluid with constant mass density, Lagrangian and Eulerian quantities are related as follows: for all suitable measurable functions g, the Reynolds average $\langle g(U)\rangle$ is defined as

$$\langle g(U)\rangle(t,x) = \mathbb{E}[g(U_t) \mid X_t = x].$$

The covariance of the velocity field, that is, its Reynolds stress tensor, is then supposed to satisfy

$$\langle \mathbf{u}^i\mathbf{u}^j\rangle = \langle U^iU^j\rangle - \langle U^i\rangle\langle U^j\rangle.$$

Assuming that (X_t, U_t) is a McKean process, the coefficients of its generator are designed such that the Lagrangian laws are consistent with closed Reynolds Average Navier-Stokes equations and other relevant physical laws. The probability distributions of the Lagrangian velocity, the pressure, etc., are suitably related to the corresponding Eulerian fields. For example, in [41] the simplified Langevin model characterizes the position X_t and velocity U_t of a fluid particle as a McKean process whose dynamics involves functions of the type $\langle g(U)\rangle(t,x)$ and therefore singular McKean interaction kernels (compared to the coefficients in the equation (2.9), the new coefficient $\langle g(U)\rangle$ is obtained by integrating, not w.r.t. the law ν_t of the solution, but w.r.t. the conditional law of some components knowing the other ones). These dynamics also involve non-smooth coefficients and wall laws at the boundary of the domain (see, e.g., Dresden and Pope [19]). A computational model of interest is thus

of the type

$$\begin{cases} X_t &= X_0 + \int_0^t U_s\, ds, \\ U_t &= U_0 - \int_0^t \dfrac{1}{\varrho} \nabla_x \mathcal{P}(s, X_s)\, ds + \int_0^t \dfrac{\varepsilon_L(s, X_s)}{k_L(s, X_s)} \left(\mathbb{E}\left[U_s \mid X_s\right] - U_s\right) ds \\ &\quad + \int_0^t \sqrt{C_0 \varepsilon_L(s, X_s)}\, dW_s + 2 \sum_{0<s\le t} \left(V(s, X_s) - U_{s^-}\right) \mathbb{I}_{\{X_s \in \Gamma\}}, \end{cases} \tag{5.1}$$

where

$$\triangle_x \mathcal{P}(s, x) = -\sum_{i,j} \partial_{ij} \mathbb{E}[U_s^i U_s^j \mid X_s = x]. \tag{5.2}$$

Pope's simulation method for his model can be interpreted as the time discretization of the stochastic interacting particle system related to the McKean process (X_t, U_t) coupled with other equations induced by physical constraints.

The analysis and discretization of (5.1) face many difficulties: the coefficients are not smooth and depend on the probability distribution of the solution in a singular way (through conditional expectations), the particles obey a specular reflection at the boundary Γ, the dynamics are coupled with the Poisson equation (5.2), and the variance of the particle system simulation is quite large.

Bossy et al. [9] established existence and uniquess of the solution to the following simplified version of Pope's model in the whole space:

$$\begin{cases} X_t = X_0 + \int_0^t U_s ds, \\ U_t = U_0 + \int_0^t B(s, X_s, U_s) ds + \sigma W_t, \\ B(s, x, u) := \mathbb{E}[b(U_s - u) \mid X_s = x], \end{cases}$$

where b is a bounded continuous function. Propagation of chaos was also established for the corresponding particle system. The proof uses estimates on the density of fundamental solutions of ultraparabolic PDEs obtained by Di Francesco and Polidoro [18].

Bossy and Jabir [7] studied the well-posedness of the simplified model enriched with the specular boundary condition at the boundary Γ of an hyperplane, which means that

$$U_t = U_0 + \int_0^t B(s, X_s, U_s) ds + \sigma W_t - 2 \sum_{0<s\le t} \left(U_{s^-} \cdot n(X_s)\right) n(X_s)\, \mathbb{I}_{\{X_s \in \Gamma\}},$$

and they proved the crucial no-permeability boundary condition

$$\mathbb{E}[U_t \cdot n(x) \mid X_t = x] = 0 \text{ a.e. in } [0, T] \times \Gamma.$$

The same authors recently extended this result to general geometries owing to a complex combination of PDE techniques for the analysis of the Fokker-Planck-McKean-Vlasov equation with specular boundary condition and stochastic calculus techniques to construct the process (X_t, U_t): see [8].

We conclude this subsection by mentioning another stochastic Lagrangian model for the Navier-Stokes equation in the whole Euclidean space: see Iyer and Mattingly [24].

5.2. Boundary conditions for Navier-Stokes equation. The vortex sheet and vortex blob methods were introduced by A. J. Chorin in a series of seminal papers (e.g., [13] and the list of references in [14]). Originally, they aim to approximate the Prandtl equation for turbulent

flows in boundary layers by means of a stochastic grid free numerical method. Interacting particles have dynamics of the type (2.8); their interaction kernel is the singular Biot and Savart kernel. In order to satisfy the no-slip condition at the boundary, artificial vorticity elements are created and added to the particle system.

Similar approaches have been developed in various directions to take more and more physics into account. For example, Goodman and Long (see references in Long [33]) have obtained convergence results for simplified models. Jourdain and Méléard [25] have proved the propagation of chaos of a particle system and established a stochastic representation for the vorticity solution to the Navier–Stokes equation with a simplified Neumann boundary condition. Benachour et al. [5] have constructed a random vortex method for the 2D Navier–Stokes equation for the vorticity by interpreting the no-slip boundary condition in terms of births or deaths of the particles of a non-linear branching diffusion process. Constantin and Iyer [15] have constructed another stochastic representation of Navier–Stokes equations with no-slip condition at the boundary of a domain, which might be the key tool to interpret and analyze the random vortex methods for boundary layers.

The convergence rate analysis of efficient simulation methods derived from the preceding representations is an open issue. To give an example of the difficulties to overcome, let us briefly comment on the stochastic representation obtained in [5] for the vorticity $\omega(t,x)$ under the constraint of the no-slip condition for the flow velocity at the boundary: D being the domain in $\mathbb{R}^2$ in which the flow is confined, for all bounded Borel functions h defined on D, one has

$$\int_D h(x)\omega(t,x)dx = \mathbb{E}\left[h(X_t)\exp\left(\int_0^t \phi\left(\omega\left(s,X_s\right)\right)dA_s\right)\right].$$

Here, ϕ is a non-signed non-linear function of the vorticity, (A_t) is the local time of X at the boundary of D, and

$$X_t = X_0 - \int_0^t (\nabla^\perp G * \omega)(s,X_s)ds - \int_0^t n(X_s)dA_s,$$

where G is the Green function for the Laplace operator in D, $n(x)$ is the unit outer normal vector to the boundary, and $\omega(t,x)$ is the probability density of the process (X_t) modified by the multiplicative functional

$$\exp\left(\int_0^t \phi\left(\omega\left(s,X_s\right)\right)dA_s\right).$$

This probability density is proven to solve the 2D Navier Stokes equation for the vorticity. As proposed by the authors, the corresponding particle system would interact by means of destruction or birth of particles at each time one of them hits the boundary; the complex rule to create or kill particles is expressed in terms of the vorticity which is approximated by means of the empirical distribution of the particles.

5.3. A model in population dynamics. Another interesting situation where particle interactions are governed by geometric rule, and for which a full numerical analysis is an open problem, was recently tackled by Villemonais [45] (see also references therein). The motivation comes from the study of Yaglom limits of biological populations.

Consider the particle system

$$X_t^{(i)} = X_0^{(i)} - \int_0^t q_i^N(X_s^{(i)})ds + W_t^{(i)},\ 1 \le i \le N,$$

where the coefficients q_i^N are locally Lipschitz. The particles start independently in the domain D which here may be unbounded, and are absorbed at Γ. It is easy to prove that, almost surely, two particles cannot be absorbed at the same time.

For all $(x_1, \ldots, x_N)$ such that one of the x_i belongs to Γ, we are given a jump measure $\mathcal{J}(x_1, \ldots, x_N)$ supported by D. At each time one particle hits Γ it jumps to a new position inside D: more precisely, if the particle i hits Γ at time τ, then its new position in D is sampled according to a jump measure $\mathcal{J}(X_\tau^{(1)}, \ldots, X_\tau^{(N)})$.

Under fairly general assumptions on the functions q_i^N and the collection of jump measures $\mathcal{J}(x_1, \ldots, x_N)$ the particle system is well defined. Numerical experiments show that its simulation allows one to achieve accurate numerical approximations of Yaglom limits. However the convergence rate analysis is an open issue.

6. A singular stochastic computational model in neuroscience

Consider a finite size network of N-neurons. The following model for the membrane potential $X_t^{(i)}$ of neuron i $(i = 1, \ldots, N)$ is widely admitted in the neuroscience literature:

$$X_t^{(i)} = X_0 + \int_0^t b(X_s^{(i)})ds + \frac{\alpha}{N}\sum_{j\neq i} M_t^{(j)} - M_t^{(i)} + W_t^{(i)}, \tag{6.1}$$

where $W_t^{(i)}$ are independent Brownian motions, $M_t^{(i)}$ is the number of times $X_t^{(i)}$ passes the threshold value of 1, i.e. the number of 'spikes', and $\alpha > 0$ is the strength of synaptic connection. After each spike, the membrane potential is reset below the threshold (at 0 when the particle is the only one to spike).

Delarue et al. [17] recently studied the mean field limit of this model as N tends to ∞:

$$\begin{cases} X_t = X_0 + \int_0^t b(X_s)ds + \alpha\mathbb{E}(M_t) - M_t + W_t, \\ M_t = \sum_{k\ge 1} \mathbb{I}_{[0,t]}(\tau_k), \\ \tau_k = \inf\{t > \tau_{k-1} : X_{t-} \ge 1\}, \tau_0 = 0. \end{cases} \tag{6.2}$$

Notice that M_t is the number of times X_t passes the threshold value 1 and that (6.2) is a non-trivial McKean-Vlasov equation since the dynamics depends on the probability distribution of (X_t) through the expectation of the singular functional M_t of the continuous trajectories of (X_t). The definition of a solution needs thus to be suitably formulated. For example, one may require that instantaneous firing rate has to remain finite:

$$e'(t) = \frac{d}{dt}\mathbb{E}(M_t) < \infty, \qquad \forall t > 0,$$

since otherwise the dynamics may blow-up (intuitively, in view of (6.1), a large number of neurons may fire at same time).

Set $p(t,y) = \mathbb{P}(X_t \in dy)$. Itô's formula gives the Fokker–Planck equation

$$\begin{cases} \partial_t p(t,y) + \partial_y[(b(y) + \alpha e'(t))p(t,y)] - \dfrac{1}{2}\partial^2_{yy} p(t,x) = \delta_0(y)e'(t), \ \ y < 1, \\ e'(t) = -\dfrac{1}{2}\partial_y p(t,1), \end{cases}$$

with boundary conditions $p(t,1) = p(t,-\infty) = 0$ and initial condition $p(0,y) = p_0(y)$. This non-classical non-linear PDE has been studied by Carrillo et al. [12] (see also references therein). Solutions may blow-up if $\alpha \geq 1$. In addition, for any $\alpha > 0$ there exists an initial condition X_0 such that blow-up occurs in finite time.

The stochastic approach developed in [17] provides an answer to the converse question: given an initial condition $X_0 = x_0$, can one find $\alpha > 0$ such that blow-up does not occur?

Theorem 6.1. *Suppose that the function b is globally Lipschitz. For any $\varepsilon > 0$ there exists an $\alpha_0 > 0$ such that whenever $X_0 = x_0 < 1 - \varepsilon$ and $\alpha \in (0, \alpha_0)$, there exists a unique process (X_t, M_t) which is a solution to the limit equation (6.2) on any $[0,T]$ that does not blow-up.*

Delarue et al. are now studying the propagation of chaos effect for the computational particle system (6.1) and the convergence rate of the empirical distribution to the probability distribution of (X_t). The construction and analysis of numerical methods for (6.2) are open questions.

7. Estimates for the statistical error

As noticed in Section 2 the statistical error can be estimated by using non-asymptotic versions of the central limit theorem. More accurate estimates can be derived from concentration inequalities. An important result has recently been obtained by Lemaire and Menozzi [32] under weak assumptions. To simplify the notation we here limit ourselves to the case of time homogeneous coefficients.

Theorem 7.1. *Suppose that the drift coeffcient b is bounded and that the matrix $a(x)$ satisfies the ellipticity condition (3.1). Suppose also that A is Hölder continuous. Then there exist constants c, C, α such that, for all Lipschitz functions f with Lipschitz constant less than 1,*

$$\forall r > 0, \ \forall N \geq 1, \ \mathbb{P}[|e_s(n,N)| \geq r + 2\sqrt{\alpha}\log(C)] \leq 2e^{-\frac{N}{\alpha}r^2}.$$

In addition, the constant α is explicit in terms of T, c, C, and the constant c and C are related to Gaussian lower and upper bounds for the probability density of $\overline{X}_T$. These bounds are obtained by adapting the parametrix method for fundamental solutions of parabolic PDEs.

8. Conclusion

We have summarized a few recent analytical and numerical advances related to continuous stochastic computational models with singular dynamics. A less succinct presentation

should for example include stochastic kinetic models, stochastic particle systems with coagulation, fragmentation or coalescence, branching stochastic dynamics and their various computational applications in biology and ecology, computational models for free energies, stochastic partial differential equations, etc.

All these problems are connected to important open theoretical and algorithmic questions such as sensitivity of the results to model uncertainties, variance reduction methodologies, and efficient dimension reduction methods.

References

[1] Alfonsi, A., *High order discretization schemes for the CIR process: Application to affine term structure and Heston models*, Math. Comp. **79**(269) (2010), 209–237.

[2] Antonelli, F. and Kohatsu-Higa, A., *Rate of convergence of a particle method to the solution of the McKean-Vlasov equation*, Ann. Appl. Probab. **12** (2002), 423–476.

[3] Bally, V. and Talay, D., *The law of the Euler scheme for stochastic differential equations (I) : convergence rate of the distribution function*, Probab. Theory Related Fields **104** (1996), 43–60.

[4] Bernardin, F., Bossy, M., Martinez, M., and Talay, D., *On mean discounted numbers of passage times in small balls of Itô processes observed at discrete times*, Electron. Comm. Probaba. **14** (2009), 302–316.

[5] Benachour, S., Roynette, B., and Vallois, P., *Branching process associated with 2D Navier–Stokes equation*, Rev. Iberoamericana **17**(2) (2001), 331–373.

[6] Bossy, M., *Optimal rate of convergence of a stochastic particle method solutions of 1D viscous scalar conservation laws*, Math. Comp. **73**(246) (2004), 777–812.

[7] Bossy, M. and Jabir, J-F., *On confined McKean Langevin processes satisfying the mean no-permeability boundary condition*, Stoch. Proc. Appl. **121**(12) (2011), 2751–2775.

[8] ———, *Lagrangian stochastic models with specular boundary condition*, Preprint arXiv:1304.6050 (2013), submitted.

[9] Bossy, M., Jabir, J-F., and Talay, D., *On conditional McKean Lagrangian stochastic models*, Probab. Theory Related Fields **151** (2011), 319–351.

[10] Bossy, M. and Jourdain, B., *Rate of convergence of a particle method for the solution of a 1 D viscous scalar conservation law in a bounded interval*, Ann. Probab. **30**(4) (2002), 1797–1832.

[11] Bossy, M., Champagnat, N., Maire, S., and Talay, D., *Probabilistic interpretation and random walk on spheres algorithms for the Poisson-Boltzmann equation in Molecular Dynamics*, ESAIM:M2AN **44**(5) (2010), 997–1048.

[12] Carrillo, J.A., González, M.d.M., Gualdani, M.P., and Schonbek, M.E., *Classical solutions for a nonlinear Fokker-Planck equation arising in computational neuroscience*, Comm. Partial Diff. Equations **38** (2013), 385–409.

[13] Chorin, A.J., *Vortex sheet approximation of boundary layers*, J. Comput. Phys. **27** (1978), 428–442.

[14] Chorin, A.J. and Marsden, J.E., *A Mathematical Introduction To Fluid Mechanics*, Springer-Verlag, New York, 1993.

[15] Constantin, P. and Iyer, G., *A stochastic-Lagrangian approach to the Navier–Stokes equations in domains with boundary*, Annals Appl. Probab. **21**(4) (2011), 1466–1492.

[16] Deaconu, M. and Herrmann, S., *Hitting time for Bessel processes - Walk on Moving Spheres Algorithm (WOMS)*, Annals Appl. Probab. **23**(6) (2013), 2259–2289.

[17] Delarue, F., Inglis, J., Rubenthaler, S., and Tanré, E., *Global solvability of a networked integrate-and-fire model of McKean-Vlasov type* (2012), submitted.

[18] Di Francesco, M. and Polidoro, S., *Schauder estimates, Harnack inequality and Gaussian lower bound for Kolmogorov-type operators in non-divergence form*, Adv. Differ. Equations **11**(11) (2006), 1261–1320.

[19] Dreeben, T.D. and Pope, S.B., *Wall-function treatment in PDF methods for turbulent flows*, Physics of Fluids **9** (1997), 2692–2703.

[20] Étoré, P., *On random walk simulation of one-dimensional diffusion processes with discontinuous coefficients*, Electron. J. Probab. **11**(9) (2006), 249–275.

[21] Gobet, E. and Menozzi, S., *Exact approximation rate of killed hypoelliptic diffusions using the discrete Euler scheme*, Stoch. Proc. Appl. **112**(2) (2004), 201–223.

[22] Graham, C. and Méléard, S., *Stochastic particle approximations for generalized Boltzmann models and convergence estimates*, Ann. Probab. **25**(1) (1997), 115–132.

[23] Graham, C. and Talay, D., *Stochastic Simulation And Monte Carlo Methods. Mathematical Foundations Of Stochastic Simulations*. Stochastic Modeling and Applied Probability Series 68, Springer, 2013.

[24] Iyer, G. and Mattingly, J., *A stochastic-Lagrangian particle system for the Navier-Stokes equations*, Nonlinearity **21**(11) (2008), 2537–2553.

[25] Jourdain, B. and Méléard, S., *Probabilistic interpretation and particle method for vortex equations with Neumann's boundary conditions*, Proc. Edimb. Math. Soc. **47**(3) (2004), 597–624.

[26] Kusuoka, S. and Stroock, D. *Applications of the Malliavin Calculus, part II*, J. Fac. Sci. Univ. Tokyo **32** (1985), 1–76.

[27] Kohatsu-Higa, A., *Weak approximations: A Malliavin calculus approach*, Math. Comp. **70** (2001), 135–172.

[28] Kohatsu-Higa, A., Lejay, A., and Yasuda, K., *Weak approximation errors for stochastic differential with non-regular coefficients*, submitted (2013).

[29] Le Gall, J-F. *One-dimensional stochastic differential equations involving the local times o f the unknown process*, In Proceedings Stochastic Analysis and Applications (Swansea, 1983), Lecture Notes in Math. **1095**, Springer, 51–82, 1984.

[30] Lejay, A., *On the constructions of the Skew Brownian motion*, Probab. Surv. **3** (2006), 413–466.

[31] Lejay, A. and Martinez, M., *A scheme for simulating one-dimensional diffusions with discontinuous coefficients*, Ann. Appl. Probab. **16**(1) (2006), 107–139.

[32] Lemaire, V. and Menozzi, S., *On some non asymptotic bounds for the Euler scheme*, Electronic J. Probab. **15** (2010), 1645–1681.

[33] Long, D.G., *Convergence of the random vortex method in two dimensions*, J. Am. Math. Soc. **1**(4) (1988), 779–804.

[34] Martinez, M. and Talay, D., *One-Dimensional parabolic diffraction equations: Pointwise estimates and discretization of related stochastic differential equations with weighted local times*, Electronic Journal Probab. **17**(27) (2012), 1–30.

[35] Mascagni, M. and Simonov, N.A., *Monte Carlo methods for calculating some physical properties of large molecules*, SIAM J. Sci. Comput. **26**(1) (2004), 339–357.

[36] Méléard, S., *A trajectorial proof of the vortex method for the two-dimensional Navier-Stokes equation*, Ann. Appl. Probab. **10**-4 (2000), 1197–1211.

[37] Milstein G.N. and Tretyakov, M.V., *Stochastic Numerics For Mathematical Physics,* Springer Verlag, 2004.

[38] Niklitschek-Soto, S. and Talay, D., *Stochastic analysis of general interface parabolic problem* (2014) in preparation.

[39] Osada, H., *Propagation of chaos for the two dimensional Navier-Stokes equation*, In K. Itô and N. Ikeda, editors, Probabilistic Methods in Mathematical Physics, 303-334, Academic Press, 1987.

[40] Petrov, V.V. *Sums Of Independent Random Variables.* Springer-Verlag, 1975.

[41] Pope, S.B. *Turbulent Flows*. Cambridge Univ. Press, 2003.

[42] Revuz, D. and Yor, M. *Continuous Martingales And Brownian Motion.* Springer-Verlag, Berlin, 1999.

[43] Sznitman, A-S. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX-1989*, volume 1464 of Lecture Notes Math., Springer, 1991.

[44] Talay, D. and Tubaro, L., *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stoch. Analysis Appl. **8**(4) (1990), 94–120.

[45] Villemonais, D., *Interacting particle systems and Yaglom limit approximation of diffusions with unbounded drift*, Electronic J. Probab. **16** (2011), 1663–1692.

2004 Route des Lucioles,BP93, 06902 Sophia Antipolis cedex, France
E-mail: denis.talay@inria.fr

A review on subspace methods for nonlinear optimization

Ya-xiang Yuan

Abstract. In this paper, we review various subspace techniques that have been used in constructing numerical methods for solving nonlinear optimization problems. As large scale optimization problems are attracting more and more attention in recent years, subspace methods are getting more and more important since they do not require solving large scale subproblems in each iteration. The essential parts of a subspace method are how to construct subproblems defined in lower dimensional subspaces and how to choose the subspaces in which the subproblems are defined. Various subspace methods for unconstrained optimization, constrained optimization, nonlinear equations and nonlinear least squares, and matrix optimization problems are given respectively, and different proposals are made on how to choose the subspaces.

Mathematics Subject Classification (2010). Primary 65K05; Secondary 90C30.

Keywords. numerical methods, nonlinear optimization, subspace techniques, subproblems.

1. Introduction

Nonlinear optimization problems have the following form:

$$\min_{x\in\Re^n} \quad f(x) \tag{1.1}$$

$$\text{subject to} \quad c_i(x) = 0, \qquad i = 1, ..., m_e\,, \tag{1.2}$$

$$c_i(x) \geq 0, \qquad i = m_e + 1, ..., m\,, \tag{1.3}$$

where m and m_e are integers satisfying $m \geq m_e \geq 0$, $f(x)$ and $c_i(x)(i = 1, ..., m)$ are real functions defined in $\Re^n$ and at least one of functions $f(x)$ and $c_i(x)(i = 1, ..., m)$ is nonlinear. If there is no constraint, namely $m = m_e = 0$, problem (1.1) is called an unconstrained optimization problem, otherwise problem (1.1)-(1.3) is called a constrained optimization problem.

Numerical methods for nonlinear optimization are iterative. At the $k-$th iteration, if the current iterate point x_k is not a solution, we try to compute a "better" point x_{k+1} and continue the process so that it will stop at a solution or generate a sequence which, hopefully, converges to a solution.

There are mainly two classes of numerical methods for nonlinear optimization. One class is line search methods in which the next iterate point is obtained by searching along a search direction. Namely, we let

$$x_{k+1} = x_k + \alpha_k d_k \tag{1.4}$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

where $d_k \in \Re^n$ is a search direction and $\alpha_k > 0$ is a step-length. The other class of methods are trust region algorithms, where a trial step s_k in a trust region is computed and then the algorithm decides whether the trail step should be accepted. The trust region is normally a small neighbourhood centered at the current iterate point x_k. Generally, the search direction or the trial step are obtained by solving a subproblem which is an approximation to the original nonlinear optimization problem. Convergence results of numerical methods for nonlinear optimization are normally based on the reduction of a penalty function. For example, the step-length α_k in a line search algorithm is chosen in such a way that sufficient reduction in the penalty function is achieved. Trial steps in a trust region algorithm will be accepted if the penalty function is reduced. A penalty function can be viewed as a combined measure for the two tasks of nonlinear optimization: reducing the objective function and satisfying the constraints. Another approach for ensuring global convergence of numerical methods for nonlinear optimization is the filter technique, which measures the constraint violation and objective function value as a two dimensional array. Detailed discussions on numerical methods for nonlinear optimization can be found in [32].

Due to their broad applications in many fields, large scale optimization problems are attracting more and more attention in recent years. However, even though the subproblems for computing search directions and trial steps are simpler than the original nonlinear optimization problems, they are still linear or quadratic problems large-scale in nature, as they are also defined in the same dimensional space as the original nonlinear problem. For example, in the $k-$th iteration, the sequential quadratic programming method for nonlinear optimization needs to solve the following quadratic programming subproblem:

$$\min_{x\in\Re^n} \quad Q_k(d) \tag{1.5}$$

$$\text{s. t.} \quad c_i(x_k) + d^T\nabla c_i(x_k) = 0, \qquad i = 1, ..., m_e\,, \tag{1.6}$$

$$c_i(x_k) + d^T\nabla c_i(x_k) \geq 0, \qquad i = m_e + 1, ..., m\,, \tag{1.7}$$

where $Q_k(d)$ is a quadratic approximation to the Lagrangian function. Though the above quadratic programming subproblem is simpler than the original nonlinear optimization problem, it is still large scale when the original nonlinear problem is large scale.

Therefore, it is important to study subspace techniques [9, 17, 41] due to the fact that subspace methods do not need to solve large scale subproblems in each iteration. In general, a subspace method searches in a lower dimensional subspace to obtain the search direction or the trust region step. Thus, in each iteration, we only need to solve a subproblem that is defined in a lower dimensional subspace.

In addition to the practical computation considerations, there are other reasons that motivated us to study numerical methods based on subspace techniques. First, let us consider a standard full space line search method. The search direction d_k is normally obtained by solving an approximation model based on the full space. For example, the search direction of the Newton's method is obtained by minimizing the second order Taylor expansion of a general nonlinear function in the whole space. Therefore, one can view that the computation of d_k is very aggressive as it is obtained through an optimistic approach by trusting the corresponding approximate model in the whole space. Once d_k is obtained, the line search procedure of computing the step-length α_k tries to minimize the one dimensional function $f(x_k + \alpha d_k)$. Thus, the computation of α_k is very conservative as it is obtained by searching in a one dimensional subspace. Thus, a standard full space line search algorithm swings between full space approximations and one-dimensional subspace searches.

Another motivation is from our long time studies on nonlinear conjugate gradient methods [10]. The search direction of a nonlinear conjugate gradient method for unconstrained optimization problem (1.1) has the form

$$d_k = -\nabla f(x_k) + \beta_k d_{k-1}\,, \tag{1.8}$$

where β_k is defined by certain conjugate conditions. Typical choices of β_k are as follows:

$$\beta_k^{HS} = \frac{g_{k+1}^T(g_{k+1}-g_k)}{d_k^T(g_{k+1}-g_k)}\,, \qquad \beta_k^{FR} = \frac{||g_{k+1}||_2^2}{||g_k||_2^2}\,, \tag{1.9}$$

$$\beta_k^{PRP} = \frac{g_{k+1}^T(g_{k+1}-g_k)}{||g_k||_2^2}\,, \qquad \beta_k^{DY} = \frac{||g_{k+1}||_2^2}{d_k^T(g_{k+1}-g_k)}\,. \tag{1.10}$$

We have two observations on the nonlinear conjugate gradient methods. Firstly, no matter which β_k is used, the new point $x_{k+1} = x_k + \alpha_k d_k$ is always in the 2-dimensional subspace $x_k + \text{span}\{-g_k, d_{k-1}\}$. Secondly, the conjugacy property is a good property only when it is associated with exact line searches. Therefore, instead of studying which formulae for β_k would lead to a good nonlinear conjugate gradient method, we should ask ourselves a different question: which point x in the two-dimensional space $x_k + \text{span}\{-g_k, d_{k-1}\}$ is the *best* point?

The third motivation for us to study subspace algorithms is the famous limited memory quasi-Newton method. Quasi-Newton methods for nonlinear optimization use quadratic models in which the Hessian is a quasi-Newton matrix updated from iteration to iteration and satisfies the following quasi-Newton equation:

$$B_k s_{k-1} = y_{k-1}\,, \tag{1.11}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1})$. An example of quasi-Newton update is the famous Broyden-Fletcher-Goldfarb-Shanno (BFGS) update:

$$\begin{aligned} B_k &= U^{BFGS}(B_{k-1}, s_{k-1}, y_{k-1}) \\ &= B_{k-1} - \frac{B_{k-1}s_{k-1}s_{k-1}^T B_{k-1}}{s_{k-1}^T B_{k-1} s_{k-1}} + \frac{y_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}}\,. \end{aligned} \tag{1.12}$$

For extremely large scale optimization problems, such as those derived from numerical weather prediction and data assimilation, we can not afford to store a full quasi-Newton matrix. To overcome such difficulties, Liu and Nocedal[21] proposed the limited memory BFGS method, which generates the quasi-Newton matrix by using the vectors s and y in the previous m iterations. Namely, $B_k^{(0)} = \sigma_k I$ and

$$B_k^{(i)} = U^{BFGS}(B_k^{(i-1)}, s_{k-m-1+i}, y_{k-m-1+i})\,,$$

for $i = 1, ..., m$. Eventually, the quasi-Newton matrix in the limited memory BFGS method has the following representation:

$$B_k = B_k^{(m)} = \sigma_k I + [S_k \quad Y_k]\, T_k \begin{bmatrix} S_k^T \\ Y_k^T \end{bmatrix}\,,$$

where T_k is a $2m \times 2m$ symmetric matrix and

$$[S_k \quad Y_k] = [s_{k-1}, s_{k-2}, ...s_{k-m}, y_{k-1}, y_{k-2}, ..., y_{k-m}] \in \Re^{n\times 2m}\,.$$

In a line search method we have $s_k = \alpha_k d_k = -\alpha_k B_k^{-1} g_k$ for some $\alpha_k > 0$, while in a trust region algorithm $s_k = -(B_k + \lambda_k I)^{-1} g_k$ for some $\lambda_k \geq 0$. Thus, in either case, we have

$$x_{k+1} - x_k \in \text{span}\{g_k, s_{k-1}, ..., s_{k-m}, y_{k-1}, ..., y_{k-m}\} . \tag{1.13}$$

This shows that limited memory quasi-Newton methods always produce a step in a lower dimensional subspace.

The block coordinate descent (BCD) method is a technique that is widely used in computational mathematics. From subspace point of view, the BCD method is a very special subspace method whose subspaces are spanned by coordinate directions. The method partitions the variables into a few blocks and then minimizes the objective function with respect to each block by fixing all other blocks at each iteration. It has been studied in convex programming [25], nonlinear programming [2], semidefinite programming [35], compressive sensing [11, 24], etc. A popular extension of the BCD method is the alternating direction method of multipliers (ADMM) by minimizing the augmented Lagrangian function blocks by blocks and then updating the Lagrangian multipliers. It dates back to optimization problems arising from partial differential equations (PDEs) [14–16], and has been applied to semidefinite programming [37], compressive sensing [40], distributed computation [5] and many other areas.

Parallel computation methods can also be viewed as subspace techniques. For example, the domain decomposition technique of Tai and Xu[39] decomposes the n dimensional space into p lower dimensional subspaces using the domain decomposition technique, and p processors search in parallel in the corresponding subspaces.

A general subspace approach requires

$$x_{k+1} - x_k \in \mathcal{S}_k , \tag{1.14}$$

where $\mathcal{S}_k$ is a subspace in $\Re^n$ with the good feature that the dimension τ_k of S_k being much less than n. An advantage of subspace approaches is that the subproblems for computing searching directions or trust region trial steps are defined in lower dimensional subspaces, which enables us to solve the corresponding subproblems quickly. Moreover, for many cases, we could show that subspace approaches attain good theoretical properties as full space models.

In a subspace method, the dimension of the subspace τ_k is either fixed or updated from iteration to iteration. $\mathcal{S}_{k+1}$ is normally updated from $\mathcal{S}_k$. Often $\mathcal{S}_{k+1}$ is obtained by adding some new directions $d_i^{(k)} (i = 1, ..., m)$:

$$\mathcal{S}_{k+1} = \text{span}\{\mathcal{S}_k, d_1^{(k)}, ..., d_m^{(k)}\} .$$

The directions $d_i^{(k)}$ to be added can be randomly generated or constructed based on the iteration information at the current iterate in order to improve the subspace. Sometimes, it is reasonable to remove some directions from the current subspace to avoid redundancy or to prevent the dimension of the subspace from increasing too rapidly. Moreover, it is reasonable for us to delete directions along which significant function reductions are not possible to obtain.

2. Subspace algorithms for unconstrained optimization

Consider a trust region algorithm for unconstrained optimization

$$\min_{x\in\Re^n} f(x)\,. \tag{2.1}$$

The trust region subproblem (TRS) is normally

$$\min_{d\in\Re^n} \quad Q_k(d) = g_k^T d + \frac{1}{2} d^T B_k d \tag{2.2}$$

$$\text{s. t.} \quad \|d\|_2 \le \Delta_k\,, \tag{2.3}$$

where $g_k = \nabla f(x_k)$, B_k is an approximate to $\nabla^2 f(x_k)$ and $\Delta_k > 0$ is the trust region bound.

When the approximate Hessian B_k is generated by quasi-Newton updates, the trust region subproblem has subspace properties. First, we have the following result

Lemma 2.1 ([34]). *Suppose $B_1 = \sigma I,\ \sigma > 0$. The matrix updating formula is any one chosen from amongst SR1, PSB and Broyden family, and B_k is the k-th updated matrix. s_k is the solution of TRS, $x_{k+1} = x_k + s_k$, $g_k = \nabla f(x_k)$. Let $\mathcal{G}_k = span\{g_1, g_2, \cdots, g_k\}$. Then $s_k \in \mathcal{G}_k$ and for any $z \in \mathcal{G}_k$, $w \in \mathcal{G}_k^{\perp}$, we have*

$$B_k z \in \mathcal{G}_k, \quad B_k u = \sigma u\,. \tag{2.4}$$

The above lemma shows that quasi-Newton matrices have very nice subspace properties. Similar results for line search QN methods are given by Gill and Leonard[13].

From the above lemma, it is not difficult to prove the following theorem.

Theorem 2.2 ([34]). *If $\mathcal{S}_k = span\{g(x_1), ..., g(x_k)\}$. The subspace trust region algorithm will generate the same sequences as the full space trust region quasi-Newton algorithm for unconstrained optimization if the $B_1 = \sigma I$ and B_k is updated by SR1, PSB and Broyden's family.*

Based on the above results, a subspace trust region quasi-Newton method for large scale unconstrained optimization is presented by Wang and Yuan[34].

Now, we discuss a special trust region subproblem which makes good use of subspace properties. If we replace the $||.||_2$ by a general norm $||.||_W$ in (2.3), we obtain a general TRS subproblem

$$\min_{s\in\Re^n} \quad g^T s + \frac{1}{2} s^T B s \tag{2.5}$$

$$\text{s. t.} \quad \|s\|_W \le \Delta\,, \tag{2.6}$$

where $\|.\|_W$ is any norm in $\Re^n$. A natural question is which norm $||.||_W$ we should use. Intuitively, we should choose the norm $\|.\|_W$ properly so that the trust region subproblem can easily be solved by using the corresponding subspace properties of the objective function $g^T s + \frac{1}{2} s^T B s$. Assume that B is a limited memory quasi-Newton matrix which is expressed as $B = \sigma I + PDP^T$, where $P \in \Re^{n\times l}$ satisfies $P^T P = I$. If we define a cylinder norm:

$$\|s\|_P = \max\{\|P^T s\|_\infty,\ \|P_\perp^T s\|_2\}\,, \tag{2.7}$$

where $P_\perp^T$ is the projection onto the space orthogonal to range(P). Due to the definition of $||.||_P$, the solution s of the P norm trust region subproblem

$$\min_{s\in\Re^n} \quad g^T s + \frac{1}{2} s^T B s \tag{2.8}$$

$$\text{s. t.} \quad \|s\|_P \le \Delta\,, \tag{2.9}$$

can be expressed by $Ps_1 + P_\perp s_2$, where s_1 is the solution of the bound-constrained quadratic programming problem

$$\min_{s\in\Re^l} \quad s^T(P^T g) + \frac{1}{2} s^T(\sigma I + D)s \tag{2.10}$$

$$\text{s. t.} \quad \|s\|_\infty \le \Delta\,, \tag{2.11}$$

and s_2 is solution of the 2-norm constrained quadratic programming problem

$$\min_{s\in\Re^{n-l}} \quad s^T(P_\perp^T g) + \frac{1}{2}\sigma s^T s \tag{2.12}$$

$$\text{s. t.} \quad \|s\|_2 \le \Delta\,. \tag{2.13}$$

It is easy to see that both s_1 and s_2 have closed form solutions:

$$(s_1)_i = \begin{cases} \frac{-(P^T g)_i}{\sigma + D_{ii}} & \text{if} \quad |(P^T g)_i| < (\sigma + D_{ii})\Delta\,, \\ \Delta \text{sign}(-(P^T g)_i) & \text{otherwise}\,, \end{cases} \tag{2.14}$$

$i = 1, ..., l$, and

$$s_2 = -\min\left(\frac{1}{\sigma}\,,\ \frac{\Delta}{||P_\perp^T g||}\right) P_\perp^T g\,. \tag{2.15}$$

Numerical results based on a trust region algorithm that uses the P-norm trust region subproblem are given by [6].

In a general line search type subspace algorithm for unconstrained optimization, we obtain the search direction by solving a subproblem defined in the subspace:

$$\min_{d\in\mathcal{S}_k} m_k(d)\,, \tag{2.16}$$

where $m_k(d)$ is an approximation to $f(x_k + d)$ for d in the subspace $\mathcal{S}_k$ It would be desirable that the approximation model $m_k(d)$ has the following properties: it is easy to minimize in the subspace $\mathcal{S}_k$, it is a good approximation to f and the solution of the subspace subproblem will yield a sufficient reduction in the original objective function f.

It is natural to use quadratic approximations to the objective function. This leads to quadratic models in subspaces. Let $dim(\mathcal{S}_k) = \tau_k$ and

$$\mathcal{S}_k = \text{span}\{p_1, p_2, ..., p_{\tau_k}\}\,.$$

Define $P_k = [p_1, p_2, ..., p_{\tau_k}]$. Thus, the subspace condition $d \in \mathcal{S}_k$ is satisfied if we let $d = P_k\bar{d}$ for $\bar{d} \in \Re^{\tau_k}$. The quadratic function $Q_k(d)$ defined in the subspace can be expressed as a function $\bar{Q}_k$ in a lower dimension space $\Re^{\tau_k}$: $Q_k(d) = \bar{Q}_k(\bar{d})$.

Now, we discuss possible choices for the subspace $\mathcal{S}_k$. First, we consider the special subspace

$$\mathcal{S}_k = \text{span}\{-g_k, s_{k-1}, ..., s_{k-m}\}\,. \tag{2.17}$$

In this case, any vector d in the subspace $\mathcal{S}_k$ has the following form:

$$d = \alpha g_k + \sum_{i=1}^{m} \beta_i s_{k-i} = (-g_k, s_{k-1}, \cdots, s_{k-m})\bar{d} \tag{2.18}$$

where $\bar{d} = (\alpha, \beta_1, \cdots, \beta_m)^T \in \Re^{m+1}$. By using the secant conditions, we estimate all the second order terms of the Taylor expansion of $f(x_k + d)$ in the subspace $\mathcal{S}_k$

$$s_{k-i}^T \nabla^2 f(x_k) s_{k-j} \approx s_{k-i}^T y_{k-j}\,, \quad s_{k-i}^T \nabla^2 f(x_k) g_k \approx y_{k-i}^T g_k\,, \tag{2.19}$$

except one term $g_k^T \nabla^2 f(x_k) g_k$. Therefore, it is reasonable to use the following quadratic model in the subspace $\mathcal{S}_k$:

$$\bar{Q}_k(\bar{d}) = (-\|g_k\|^2, g_k^T s_{k-1}, \cdots, g_k^T s_{k-m})\bar{d} + \frac{1}{2}\bar{d}^T \bar{B}_k \bar{d}\,, \tag{2.20}$$

where

$$\bar{B}_k = \begin{pmatrix} \rho_k & -g_k^T y_{k-1} & \cdots & -g_k^T y_{k-m} \\ -g_k^T y_{k-1} & y_{k-1}^T s_{k-1} & \cdots & y_{k-m}^T s_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ -g_k^T y_{y-m} & y_{k-m}^T s_{k-1} & \cdots & y_{k-m}^T s_{k-m} \end{pmatrix} \tag{2.21}$$

with $\rho_k \approx g_k^T \nabla^2 f(x_k) g_k$. Hence, once we have a good estimate to the term $g_k^T \nabla^2 f(x_k) g_k$, we obtain a good quadratic model in the subspace $\mathcal{S}_k$.

There are different ways to choose ρ_k. Similarly to Stoer and Yuan[31], we let

$$\rho_k = 2\frac{(s_{k-1}^T g_k)^2}{s_{k-1}^T y_{k-1}}\,, \tag{2.22}$$

due to the fact that the mean value of $\cos^2(\theta)$ is $\frac{1}{2}$, which gives

$$g_k^T \nabla^2 f(x_k) g_k = \frac{1}{\cos^2\theta_k}\frac{(s_{k-1}^T \nabla^2 f(x_k) g_k)^2}{s_{k-1}^T \nabla^2 f(x_k) s_{k-1}} \approx 2\frac{(s_{k-1}^T g_k)^2}{s_{k-1}^T y_{k-1}}\,, \tag{2.23}$$

where θ_k is the angle between $(\nabla^2 f(x_k))^{\frac{1}{2}} g_k$ and $(\nabla^2 f(x_k))^{\frac{1}{2}} s_{k-1}$. Another way to estimate $g_k^T(\nabla^2 f(x_k))g_k$ is to replace $\nabla^2 f(x_k)$ by a quasi-Newton matrix. We can also obtain ρ_k by computing an extra function value $f(x_k + tg_k)$ and setting

$$\rho_k = \frac{2(f(x_k + tg_k) - f(x_k) - t\|g_k\|_2^2)}{t^2}\,. \tag{2.24}$$

By letting the second order curvature along g_k to be the average of those along s_{k-i} $(i = 1, ..., m)$, we get

$$\rho_k = \frac{\|g_k\|_2^2}{m}\sum_{i=1}^{m}\frac{s_{k-i}^T y_{k-i}}{s_{k-i}^T s_{k-i}}\,. \tag{2.25}$$

Suppose $g_k^T \nabla^2 f(x_k) g_k = \rho$, we have $d(\rho) =$

$$(-g_k, s_{k-1}, \ldots, s_{k-m}) \begin{pmatrix} \rho & -g_k^T y_{k-1} & \cdots & -g_k^T y_{k-m} \\ -g_k^T y_{k-1} & y_{k-1}^T s_{k-1} & \cdots & y_{k-m}^T s_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ -g_k^T y_{y-m} & y_{k-m}^T s_{k-1} & \cdots & y_{k-m}^T s_{k-m} \end{pmatrix}^{-1} \begin{pmatrix} -\|g_k\|^2 \\ g_k^T s_{k-1} \\ \cdots \\ g_k^T s_{k-m} \end{pmatrix}$$

Using

$$(B + \rho e e^T)^{-1} = B^{-1} - \frac{\rho}{1 + \rho e^T B^{-1} e} B^{-1} e e^T B \,,$$

we could show that the solution set is on a line:

$$d(\rho) = d(+\infty) + \alpha(\rho) \hat{d} \,.$$

Thus, instead of estimating an ideal ρ, we can carry out a line search for ρ to achieve sufficient reduction in the objective function.

Similar to (2.17), a slightly different subspace is

$$\mathcal{S}_k = \text{span}\{-g_k, y_{k-1}, ..., y_{k-m}\} \,. \tag{2.26}$$

In this case, any vector in $\mathcal{S}_k$ is represented as

$$d = \alpha g_k + \sum_{i=1}^{m} \beta_i y_{k-i} = W_k \bar{d} \tag{2.27}$$

where $W_k = [-g_k, y_{k-1}, ..., y_{k-m}] \in \Re^{n \times (m+1)}$. The Newton's step in the subspace $\mathcal{S}_k$ is $W_k \bar{d}_k$ with

$$\bar{d}_k = -\left[W_k^T \nabla^2 f(x_k) W_k\right]^{-1} W_k^T \nabla f(x_k) \,. \tag{2.28}$$

Thus, the remaining issue we need to consider is to obtain a good estimate of $\bar{d}_k$, using the fact that all the elements of $\left[W_k^T (\nabla^2 f(x_k))^{-1} W_k\right]$ is known except one entry $g_k \nabla^2 f(x_k)^{-1} g_k$.

Due to the property of (1.13), it is reasonable to use

$$\mathcal{S}_k = \text{span}\{-g_k, s_{k-1}, ..., s_{k-m}, y_{k-1}, ..., y_{k-m}\} \,. \tag{2.29}$$

This subspace is used by [33] where a subspace trust region limited memory quasi-Newton method is presented.

Now, we consider subspaces spanned by coordinate directions. Such subspaces have sparsity structures. First, let us sort $|(g_k)_i|$ by the descending order

$$|(g_k)_{i_1}| \geq |(g_k)_{i_2}| \geq |(g_k)_{i_3}| \geq \cdots \,. \tag{2.30}$$

We call the subspace

$$\mathcal{S}_k = \text{span}\{e_{i_1}, e_{i_2}, ..., e_{i_\tau}\} \tag{2.31}$$

the *τ-steepest coordinates subspace*. One good property of the steepest coordinates subspace is that the steepest descent direction in the subspace is a sufficiently descent direction, namely

$$\min_{d \in \mathcal{S}_k} \frac{d^T g_k}{||d||_2 ||g_k||_2} \leq -\frac{\tau}{n} \,. \tag{2.32}$$

If $(g_k)^2_{i_{\tau+1}} \leq \epsilon \sum_{j=1}^{\tau} (g_k)^2_{i_j}$, we obtain the following estimate:

$$\min_{d \in \mathcal{S}_k} \frac{d^T g_k}{||d||_2 ||g_k||_2} \leq -\frac{1}{\sqrt{1+\epsilon(n-\tau)}}\,. \tag{2.33}$$

By sequentially adding steepest coordinate directions into the subspace, we obtain a *sequential steepest coordinates search* (SSCS) technique. As an example, let us consider applying the sequential steepest coordinates search to the minimization of a convex quadratic function

$$Q(x) = g^T x + \frac{1}{2} x^T B x\,.$$

Algorithm 2.3. (Sequential steepest coordinates search for quadratic functions)

Step 1 *Given* x_1. $k := 1$.

Step 2 *Compute* $g_k = \nabla Q(x_k)$, *if* $||g_k|| = 0$ *then stop;*
Choose $i_k = \arg\min_i \{|(g_k)_i|\}$.

Step 3 *Let* $S_k = \text{span}\{e_{i_1}, ..., e_{i_k}\}$,
Find $x_{k+1} = \arg\min_{x \in x_1 + S_k} Q(x)$;
Go to Step 2.

The sequential steepest coordinates search could be used to obtain an approximate sparse solution of linear least square problems. For example, consider the following sparsity constraint linear least squares problem:

$$\min_{x \in \Re^n} \quad ||Ax - b||_2^2 \tag{2.34}$$

$$\text{s. t.} \quad ||x||_0 \leq r\,, \tag{2.35}$$

where $A \in \Re^{n \times m}$, $b \in \Re^m$, r is a positive integer less than n, and $||x||_0$ is the number of non-zero elements of vector x. If Algorithm 2.3 is applied to $\min Q(x) = \frac{1}{2}||Ax - b||_2^2$, it will give a greedy algorithm for (2.34)-(2.35).

Algorithm 2.4. (SSCS for linear least squares)

Step 1 $x_1 = 0$, $g = A^T b$, $i_1 = \arg\max\{|(g)_i|\}$, $p_1 = e_{i_1}$, *given* $\epsilon > 0$.

Step 2 $\alpha_k = \arg\min_\alpha Q(x_k + \alpha p_k)$,
$x_{k+1} = x_k + \alpha_k p_k$,

Step 3 *If* $k \geq r$ *then stop;* $g := g - \alpha A^T A p_k$;
If $||g||_2 \leq \epsilon$ *then stop;*

Step 4 *let* $i_{k+1} = \arg\max_i \{|(g)_i|\}$;
let $p_{k+1} \in \text{span}\{p_1, ..., p_k, e_{i_{k+1}}\}$ *conjugate to* $p_1, ..., p_k$.

Step 5 $k := k + 1$, *go to Step 2.*

If $\epsilon = 0$, the solution obtained by the above algorithm is a local solution of problem (2.34)-(2.35). Let $S(r, A, b)$ be the set of all global solutions of (2.34)-(2.35), we are interested in studying what conditions would imply $x_{r+1} \in S(r, A, b)$. If $A = I$, it is easily to see that $x_{r+1} \in S(r, I, b)$. For general A, if $r = 1$ or 2 we have the following results.

Lemma 2.5. *Let $A = (a_1, ..., a_n)$. If $||a_i|| = 1$ for all i, the iterate point x_{k+1} obtained by the SSCS algorithm has the following properties:*

(1) $x_2 \in S(1, A, b)$;

(2) *There exists a $y \in S(2, A, b)$ such that x_3 and y share one non-zero element index.*

The subproblems in the SSCS algorithm have the form

$$\min_{x \in \Re^n} \quad ||Ax - b||_2^2 \tag{2.36}$$

$$\text{s. t.} \quad x_i = 0\,, \quad i \in I_k \tag{2.37}$$

for some active set I_k. Thus, general subspaces spanned by coordinate directions for sparsity constraint problems should have the form $\mathcal{S}_k = \{d \mid d_i = 0, i \in I_k\}$. Such subspaces are used by many methods for compressive sensing. One particular optimization model in compressive sensing is the l_0 minimization problem

$$\min_{x \in \Re^n} \quad ||x||_0 \tag{2.38}$$

$$\text{s. t.} \quad Ax = b\,. \tag{2.39}$$

For more detailed discussions, please refer to [38] and the references given there.

Another possible subspace is the *steepest descent $\tau-$subspace*, which is a τ dimensional subspace which forces τ elements of the gradient vector to be zero. Instead of requiring the whole vector $g(x) = 0$, which is the optimality condition for $\min f(x)$, we require τ elements of $g(x)$ to be zero, namely

$$\bar{g}(x) = ((g(x))_{i_1}, (g(x))_{i_2}, ..., (g(x))_{i_\tau})^T = 0\,,$$

at the current iteration. This should be achievable by searching in a subspace spanned by τ coordinate directions, since there are only τ equations. Let the Jacobian of $\bar{g}(x)$ to be $\bar{A}(x)$, a Newton's step d satisfies

$$(\bar{A}(x_k))^T d + \bar{g}(x_k) = 0\,. \tag{2.40}$$

Because the above system has τ equations with n unknowns, it is possible to consider d in any subspace spanned by τ coordinate directions. There are C_n^m such choices, and we call the one which makes the length of the solution of (2.40) in the subspace the shortest as the steepest descent τ-subspace. Intuitively, this subspace has the nice property of forcing τ elements of the gradient vector to zero by moving a $\tau-$coordinate step as small as possible. However, such a definition of the subspace seems to be too theoretical and may not be easy to be implemented in practice, as it needs to solve linear least squares problem with linear constraints and a sparsity constraint:

$$\begin{aligned} \min_{d \in \Re^n} \quad & ||d||_2^2 \\ \text{s. t.} \quad & (\bar{A}(x_k))^T d + \bar{g}(x_k) = 0, \quad ||d||_0 = \tau\,. \end{aligned}$$

3. Subspace techniques for constrained optimization

Now we consider subspace techniques for constrained optimization. In order to simplify the presentation, instead of considering the general problem (1.1)-(1.3), we focus on the equality

constrained problem:

$$\min_{x\in\Re^n} \quad f(x) \tag{3.1}$$

$$\text{s. t.} \quad c(x) = 0\,, \tag{3.2}$$

where $c(x) = (c_1(x), \cdots, c_m(x))^T$.

The sequential quadratic programming method (SQP) is an important numerical method for solving constrained optimization. The main idea of the SQP method is to solve the nonlinearly constrained problem (3.1)-(3.2) by successively minimizing quadratic approximations to the Lagrangian function subject to the linearized constraints. The search direction d_k of a line search type SQP method is obtained by solving the following quadratic programming subproblem

$$\min_{d\in\Re^n} \quad Q_k(d) = g_k^T d + \frac{1}{2} d^T B_k d \tag{3.3}$$

$$\text{s. t.} \quad c(x_k) + A_k^T d = 0\,, \tag{3.4}$$

where $A_k = \nabla c(x_k)$ and B_k is an approximation to the Hessian of the Lagrangian function. The SQP step d_k can be decomposed into two parts $d_k = h_k + v_k$ where $v_k \in \text{range}(A_k)$ and $h_k \in \text{null}(A_k^T)$. Thus, v_k is a solution of the linearized constrained constraints (3.4) in the range space of A_k, while h_k is the minimizer of the quadratic function $Q_k(v_k + d)$ in the null space of A_k^T.

One good property of the SQP method is that it converges superlinearly, namely when x_k is close to a KKT point x^* we have the following relation

$$x_k + d_k - x^* = o(\|x_k - x^*\|)\,. \tag{3.5}$$

But, the superlinearly convergent step d_k may lead to a point that seems "bad" as it may increase both the objective function and the constraint violations. The famous Marotos effect shows that it is possible for the SQP step d_k to have both $f(x_k + d_k) > f(x_k)$ and $\|c(x_k + d_k)\| > \|c(x_k)\|$, even though (3.5) holds. A remedy for overcoming the Marotos effect is the second order correction step method[12, 26], where the step is obtained by resolving the quadratic programming subproblem with the constraints (3.4) are replaced by

$$c(x_k + d_k) + A_k^T(d - d_k) = 0 \tag{3.6}$$

because the left hand side of (3.6) is a better approximation to $c(x_k + d)$ near the point $d = d_k$. Since the change of the constraints is a second order term, the new step can be viewed as the SQP step d_k adding a second order correction step $\hat{d}_k$. For detailed discussions on the SQP method and the second order correction step, please see [32].

Now, let us examine the second order correction step from subspace point of views. The second order correction step $\hat{d}_k$ is a solution of

$$\min_{d\in\Re^n} \quad Q_k(d_k + d) \tag{3.7}$$

$$\text{s. t.} \quad c(x_k + d_k) + A_k^T d = 0\,. \tag{3.8}$$

Assume that the QR factorization of A_k is $[Y_k, Z_k]\begin{bmatrix} R_k \\ 0 \end{bmatrix}$ and R_k is nonsingular. Thus, the second order correction step is represented as $\hat{d}_k = \hat{v}_k + \hat{h}_k$, where $\hat{v}_k = -Y_k R_k^{-T} c(x_k + d_k)$

and $\hat{h}_k$ is the minimizer of

$$\min_{h \in \text{null}(A_k^T)} \quad Q(d_k + \hat{v}_k + h) \,. \tag{3.9}$$

Since d_k is the SQP step, it follows that $g_k + B_k d_k \in \text{range}(A_k)$, which implies that the minimization problem (3.9) is equivalent to

$$\min_{h \in \text{null}(A_k^T)} \quad \frac{1}{2}(\hat{v}_k + h)^T B_k (\hat{v}_k + h) \,. \tag{3.10}$$

If $Y_k^T B_k Z_k = 0$, we have that $\hat{h}_k = 0$, which shows that the second order correction step $\hat{d}_k \in \text{range}(A_k)$ is also a range space step. In this case, the second order correction uses two range space steps and one null space step. This is an undesirable property because a range space step is a fast convergent step as it is a Newton's step while a null space step is normally a slower convergent step due to the fact that it is normally a quasi-Newton step because B_k is generally a quasi-Newton approximation to the Hessian of the Lagrangian function. Hence, examining the SQP method with subspace properties helps us to understand the insights of the method. Intuitively, it would be more reasonable to have two steps in the slower space with one step in the fast space. Thus, it might be better to investigate a modified SQP method with a correction step $\hat{d}_k \in \text{null}(A_k^T)$.

We can also consider subspaces other than the null space and the range space. In general, a subspace SQP method obtains the search direction d_k by solving a QP in a subspace:

$$\min_{d \in \Re^n} \quad Q_k(d) \tag{3.11}$$

$$\text{s. t.} \quad c_k + A_k^T d = 0, \qquad d \in \mathcal{S}_k \,, \tag{3.12}$$

where $\mathcal{S}_k$ is a subspace. Lee[20] considered the following choice:

$$\mathcal{S}_k = \text{span}\{-g_k, d_1, ..., d_{k-1}, -\nabla c_{k_i}\} \,,$$

where $|c_{k_i}| = \|c_k\|_\infty$.

In some trust region algorithms for constrained optimization, the subproblem that needs to be solved in each iteration is the Celis-Dennis-Tapia subproblem[7]

$$\min_{d \in \Re^n} \quad Q_k(d) = g_k^T d + \frac{1}{2} d^T B_k d \tag{3.13}$$

$$\text{s. t.} \quad ||c_k + A_k^T d||_2 \le \xi_k, \quad ||d||_2 \le \Delta_k \,. \tag{3.14}$$

Recently, It is shown that the CDT subproblem has certain subspace properties[18]:

Lemma 3.1 ([18]). *Let $S_k = span\{Z_k\}$, $Z_k^T Z_k = I$, $span\{A_k, g_k\} \subset S_k$ and $B_k u = \sigma u$, $\forall u \in S_k^\perp$. Then the CDT subproblem is equivalent to*

$$\min_{\bar{d} \in \Re^r} \quad \bar{Q}_k(\bar{d}) = \bar{g}_k^T \bar{d} + \frac{1}{2} \bar{d}^T \bar{B}_k \bar{d} \tag{3.15}$$

$$s.\ t. \quad ||c_k + \bar{A}_k^T \bar{d}||_2 \le \xi_k, \quad ||\bar{d}||_2 \le \Delta_k \,, \tag{3.16}$$

where $\bar{g}_k = Z_k^T g_k$, $\bar{B}_k = Z_k^T B_k Z_k$ and $\bar{A}_k = Z_k^T A_k$.

Based on the above result, a subspace version of the Powell-Yuan trust algorithm[28] was given in [18].

Subspace techniques can also be used with other methods for constrained optimization. For example, interior methods for nonlinearly constrained optimization basically use a Newton's step to the KKT system based on the log-barrier function. If we solve the derived linear system in a lower dimensional subspace, it will give us a subspace version of an interior point method.

There are many subspace techniques for bound-constrained problems, where the constraints are

$$l \leq x \leq u\,, \tag{3.17}$$

where l and u are two given vectors in $\Re^n$. For example, A subspace adaptation of the Coleman-Li trust region and interior method[8] is proposed for solving large-scale bound-constrained minimization problems[3], and another subspace version of the Coleman-Li trust region algorithm was presented in [41]. Ni and Yuan[27] proposes a subspace limited memory quasi-Newton method for solving large-scale optimization with bound constraints (3.17), in which the limited memory quasi-Newton method is used to update the variables with indices outside of the active set, while the projected gradient method is used to update the active variables.

4. Subspace techniques for nonlinear equations and nonlinear least squares

In this subsection, we consider systems of nonlinear equations

$$F_i(x) = 0, \quad i = 1, ..., m; \quad x \in \Re^n\,, \tag{4.1}$$

and nonlinear least squares:

$$\min_{x \in \Re^n} \sum_{i=1}^{m} (F_i(x))^2\,. \tag{4.2}$$

Because nonlinear least squares problem (4.2) is a special unconstrained optimization problem, all the subspace techniques discussed in Section 2 can be applied. Due to the special structures of nonlinear equations and nonlinear least squares, there are special subspace approaches. For example, several implementations of Newton-like iteration schemes based on Krylov subspace projection methods for solving nonlinear equations are considered in [4]. The Gauss-Seidel iteration for linear equations can be extended for nonlinear equations. In the following, we will discuss some possible subspace approaches including incomplete sum, partition of variables, and steepest descent $\tau-$subspace.

First, we explain the technique of incomplete sum for nonlinear least squares. At iteration k, we minimize the sum of squares of some selected terms instead of all terms. Namely, define an index set J_k which is a subset of $\{1, ..., m\}$, and consider

$$\min_{x \in \Re^n} \sum_{i \in J_k} (F_i(x))^2\,. \tag{4.3}$$

The incomplete sum approach works quite well for certain class of problems, for example the distance geometry problem which has lots of applications including protein structure prediction, where the nonlinear least squares of all the terms would have lots of local minimizers[30].

For nonlinear equations, the incomplete approach is to ignore some equations. Instead of requiring the original system (4.1), we consider

$$F_i(x) = 0, \quad i \in J_k\,, \tag{4.4}$$

which is an incomplete set of equations. It is easy to see the incomplete approach is a subspace technique. Define the vector

$$F = \begin{pmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_m(x) \end{pmatrix} \in \Re^m\,.$$

To solve the nonlinear equations (4.1) is to find a x at which F maps to the origin. Let P_k^T be a mapping from R^m to a lower dimensional subspace, solving the reduced system

$$P_k^T F(x) = 0 \tag{4.5}$$

is exactly replacing $F = 0$ by requiring its mapping to the subspace spanned by P_k to be zero. In particular, if the columns of P_k are chosen to be coordinate vectors $\{e_i, i \in J_k\}$, we obtain the incomplete set of equations (4.4).

Now, we consider partition of variables, which is clearly a subspace technique. Let I_k be a subset of $\{1, ..., n\}$. We partition the variables into two group $x = (\bar{x}\,, \hat{x})$, where $\bar{x} = \{x_i, i \in I_k\}$ and $\hat{x} = \{x_i\,,\ i \notin I_k\}$. At the $k-$th iteration, we fix the variables $\hat{x}$ and allow $\bar{x}$ to change in order to obtain a better iterate point. To be exact, we try to solve

$$\min_{\bar{x} \in \Re^{|I_k|}} \sum_{i=1}^{m} (F_i(\bar{x}, \hat{x}_k))^2\,. \tag{4.6}$$

The above problem has fewer variables. It is easy to see that partition of variables use special subspaces that spanned by coordinate directions. An obvious generalization of partition of variables is decomposition of the space which uses subspaces spanned by any given directions. For example, assume that we have i_k vectors $\{q_1^{(k)}, q_2^{(k)}, ..., q_{i_k}^{(k)}\}$ which spans $\mathcal{S}_k$. Similar to (4.6), we consider the subspace subproblem

$$\min_{d \in \mathcal{S}_k} \sum_{i=1}^{m} (F_i(x_k + d))^2\,. \tag{4.7}$$

When the above subproblem is combined with the reduced system technique, it gives the general subspace subproblem for nonlinear least squares

$$\min_{d \in \mathcal{S}_k} ||P_k^T F(x_k + d)||_2^2\,. \tag{4.8}$$

For nonlinear equations, a similar subproblem is

$$P_k^T F(x_k + Q_k z) = 0, \tag{4.9}$$

where $Q_k = [\ q_1^{(k)},\ q_2^{(k)},\ ...,\ q_{i_k}^{(k)}\ ]$ and $P_k = [\ p_1^{(k)},\ p_2^{(k)},\ ...,\ p_{i_k}^{(k)}\ ]$. Let J_k be the Jacobian of F at x_k, the linearized system for subproblem (4.9) is

$$P_k^T [\ F(x_k) + J_k Q_k z\ ] = 0\,. \tag{4.10}$$

Of course, the efficiency of such an approach depends on how to select P_k and Q_k. We can borrow ideas from subspace techniques for large scale linear systems[29]. Instead of using (4.10), we construct a subproblem of the following form:

$$P_k^T F(x_k) + \hat{J}_k z = 0\,, \tag{4.11}$$

where $\hat{J}_k \in \Re^{i_k \times i_k}$ is an approximation to $P_k^T J_k Q_k$. The reason for preferring (4.11) over (4.10) is that in (4.11) we do not need the Jacobian matrix J_k, whose size is normally significantly larger than that of $\hat{J}_k$.

The $\tau-$steepest descent coordinate subspace discussed in Section 2 can also be extended to nonlinear equations and nonlinear least squares. Here we only discuss nonlinear equations. Assume we have

$$|F_{i_1}(x_k)| > \cdots > |F_{i_\tau}(x_k)| > \cdots \tag{4.12}$$

at the $k-$th iteration. A direct extension of the $\tau-$ steepest descent coordinate subspace method discussed in Section 2 would solve

$$F_{i_j}(x_k) + d^T \nabla F_{i_j}(x_k) = 0 \quad j = 1, ..., \tau\,. \tag{4.13}$$

in the subspace spanned by the corresponding coordinate directions $\{e_{i_j}, \quad j = 1, ..., \tau\}$. This approach is reasonable if $F(x)$ is a monotone operator. For general nonlinear functions $F(x)$, it seems that we should replace e_{i_j} by the coordinate direction which is the steepest descent coordinate direction of the function $F_{i_j}(x)$ at x_k. Namely, we should replace i_j by an index l_j such that

$$l_j = \text{argmax}_{t=1,...,n} \left| \frac{\partial F_{i_j}(x_k)}{\partial (x)_t} \right| .$$

However, such a choice may lead to one l_j for two different j, which makes subproblem (4.13) has no solution in the subspace spanned by $\{e_{l_1}, ..., e_{l_\tau}\}$.

A good subspace spanned by $\tau-$ coordinate directions might be the steepest descent $\tau-$subspace as discussed in Section 2, which should contain the shortest vector d from all solutions of (4.13) satisfying $||d||_0 = \tau$. However, such a subspace is not easy to obtain, an approximation could be derived by finding τ row indices of the matrix $[\nabla F_{i_1}(x_k), ..., \nabla F_{i_\tau}(x_k)]$ such that the corresponding $\tau \times \tau$ sub-matrix Γ_k makes $||(\Gamma_k)^{-1}||$ as small as possible.

More detailed discussions on subspace methods for nonlinear equations and nonlinear least squares are given in [42].

5. Subspace techniques for matrix optimization

Matrix optimization problems have stimulated lots of researches in recent years due to their broad applications. The one million dollar Netflix prize problem[1] may be formulated as the following problem

$$\min_{X \in \Re^{n \times m}} \quad \text{rank}(X) \tag{5.1}$$

$$\text{s. t.} \quad (X)_{ij} = M_{ij}\,, \quad (i,j) \in \mathcal{T}\,, \tag{5.2}$$

where $\mathcal{T}$ is a subset of $\{(i,j) \mid i = 1, ..., n;\ j = 1, ..., m\}$, and $M_{ij}((i,j) \in \mathcal{T})$ are given data. A second example of matrix optimization problem is the semidefinite programming

problem

$$\min_{X\in\Re^{n\times n}} \quad \langle C\,,\,X\rangle \tag{5.3}$$

$$\text{s. t.} \quad \langle A_i,\,X\rangle \geq b_i, \quad i=1,...,m, \tag{5.4}$$

$$X \succeq 0, \tag{5.5}$$

where $\langle X\,,\,Y\rangle = trace(X^TY)$. Another example is solving the Kohn-Sham equation in density functional theory from physics and quantum chemistry, where the total energy of a system needs to be minimized. This leads to the minimization of a nonlinear matrix function with orthogonality constraints:

$$\min_{X\in\Re^{n\times m}} \quad E(X) \tag{5.6}$$

$$\text{s. t.} \quad X^TX = I\,, \tag{5.7}$$

where $E(X)$ is the energy function [22, 36].

A general nonlinear matrix optimization has the following form

$$\min_{X\in\mathcal{X}} \quad f(X) \tag{5.8}$$

$$\text{s. t.} \quad c(X) = 0, \tag{5.9}$$

where $f:\Re^{n\times m}\to\Re$, $\mathcal{X}\subseteq\Re^{n\times m}$ and $c:\Re^{n\times m}\to\Re^p$. The constraints have been split into the set $\mathcal{X}$ and the general constraints $c(X)=0$ according to their structures and roles in the targeted subspace subproblems. For example, some simple constraints such as orthogonality and positive semidefiniteness can be put in $\mathcal{X}$ and the subspace subproblems still have a computable closed form solution. Specifically, for a suitably chosen subspace $\mathcal{S}_k\subset\Re^{n\times m}$, $m_k(X)\approx f(X)$ and an linear operator $\mathcal{A}_k$ such that $\mathcal{A}_k(X)\approx c(X)$ for $X\in\mathcal{S}_k$, the subspace subproblem is:

$$\min_{X\in\mathcal{S}_k\cap\mathcal{X}} \quad m_k(X) \tag{5.10}$$

$$\text{s. t.} \quad \mathcal{A}_k(X) = 0. \tag{5.11}$$

Then, a model subspace algorithm for the general matrix optimization problem (5.8)-(5.9) can be given as follows.

Algorithm 5.1. (Model subspace method for nonlinear matrix optimization)

Step 1 *Given X_1. Let $k:=1$.*

Step 2 *If X_k is a stationary point of (5.8)-(5.9) then stop.*
Choose a low-dimensional subspace $\mathcal{S}_k\subset\Re^{n\times m}$,
build an approximate model $m_k(X)\approx f(X)$ for $X\in\mathcal{S}_k$, and an linear operator $\mathcal{A}_k$ such that $\mathcal{A}_k(X)\approx c(X)$ for $X\in\mathcal{S}_k$.

Step 3 *Solve* (5.10)-(5.11) *to obtain $\hat{X}$.*

Step 4 *Choose a suitable map $h(X):\Re^{n\times m}\to\Re^{n\times m}$ to construct the next iteration: $X_{k+1}:=h(\hat{X}_k)$; Go to Step 2.*

Most of the techniques for choosing subspaces in subsection 2.1 can be extended here. For example, we can choose the subspace mainly spanned by the gradients at the first k iterations:

$$\mathcal{S}_k = \text{span}\{X_k, \nabla f(X_1), \cdots, \nabla f(X_k)\}, \tag{5.12}$$

or use the conjugate gradient type subspace

$$\mathcal{S}_k = \text{span}\{\nabla f(x_k),\ X_k,\ X_{k-1}\}. \tag{5.13}$$

There are various ways for defining subspaces when the matrix optimization problems have special structures. For example, for the low rank matrix optimization problems we can search in subspaces of low dimensional manifolds of low rank matrices. In particular, consider the following problem

$$\min_{X\in\Re^{n\times p}} \quad ||\mathcal{A}(X) - b||_2^2 \tag{5.14}$$

$$\text{s. t.} \quad \text{rank}(X) \le r\,. \tag{5.15}$$

One special subspace is

$$\mathcal{S}_k = \{X_k + Y \mid \text{rank}(Y) \le \tau\}. \tag{5.16}$$

If $\tau = 1$, we update the iterate matrix with the increment being a rank-1 matrix.

Computing the dominate singular value decomposition of a given matrix $A \in \Re^{n\times m}$ leads to a matrix optimization problem with orthogonality constraints:

$$\max_{X\in\Re^{n\times p}} \quad ||A^T X||_F^2 \tag{5.17}$$

$$\text{s. t.} \quad X^T X = I. \tag{5.18}$$

Let $\mathcal{X} = \{X \in \Re^{n\times p} \mid X^T X = I\}$ and $c(X) = \emptyset$. The locally optimal block preconditioned conjugate gradient method (LOBPCG) [19] chooses $h(\cdot)$ as the identity map and the following conjugate gradient type of subspace:

$$\mathcal{S}_k = \text{span}\{X_{k-1}, X_k, AA^T X_k\}, \tag{5.19}$$

The corresponding subspace problem is a $3p$-dimensional generalized eigenvalue problem which can be solved fast due to the fact that $p \ll \min\{n, m\}$. The limited memory block Krylov subspace optimization method (LMSVD, [23]) selects the subspace

$$\mathcal{S}_k = \text{span}\{X_k, X_{k-1}, ..., X_{k-q}\} \tag{5.20}$$

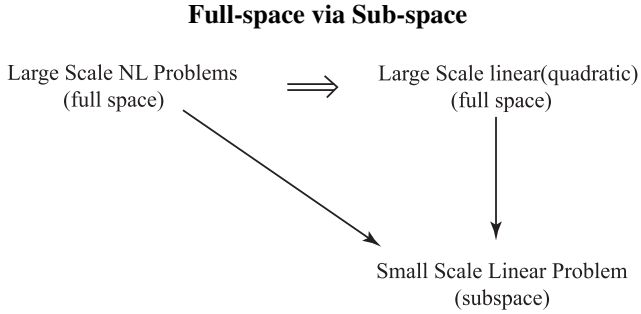
with an adaptive way to adjust the size of $\mathcal{S}_k$ and takes

$$h(X) := \text{orth}(AA^T X), \tag{5.21}$$

which reduces the probability to be trapped by saddle points of (5.17)-(5.18). A general global convergence analysis for both LOBPCG and LMSVD is established in [23] by requiring some minimal assumptions.

6. Summary

In this paper, we review subspace techniques for nonlinear optimization. Compared to full space algorithms which normally convert nonlinear problems to linear/quadratic systems without reducing the size of the problem, subspace algorithms aim to to take a short-cut from large scale nonlinear problem to small scale linear/quadratic systems. This is illustrated by the following diagram:



Subspace techniques are suitable for problems where function values are difficult to compute and problems that are highly nonlinear for which normally line searches are very expensive. Though we have given quite a few suggestions on how to choose subspaces, there are still many issues to be investigated further, including how to balance between null space and range space for constrained optimization for null-space type methods and how to choose subspaces depending on constraints for general subspace methods for constrained optimization.

The subspace techniques discussed in the paper show that large scale problems can be approximated by lower dimensional subspace subproblems, and we believe that the nice properties of subspace techniques will enable them to play an important role in the development of numerical methods for large scale optimization.

Acknowledgements. The author's research is partially supported by NSFC grants 11331012 and 11321061. I am very grateful to my former students Xin Liu, Zaiwen Wen and Cong Sun for their helpful comments on the first draft of the paper. I also thank my daughter Yuan Yuan for her polishing the English of the paper.

References

[1] J. Bennett and S. Lanning, *The Netflix prize*, Proceedings of KDD Cup and Workshop, 2007.

[2] D.P. Bertsekas. *Nonlinear Programming*, Athena Scientific, Belmont, Massachusetts, 1999.

[3] M.A. Branch, T.F. Coleman, and Y.Y. Li, *A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems*, SIAM J. Sci. Comput. **21** (2007), 1–23.

[4] P. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM Journal on Scientific and Statistical Computing **11** (1990), 450–481.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning **3** (2011), 1–122

[6] O. Burdakov, L.J. Gong, Y.X. Yuan, and S. Zikrin, *On efficiently combining limited memory and trust-region techniques*, Report, AMSS, CAS, 2013.

[7] Celis, M.R., Dennis, J.E., and Tapia, R.A., *A trust region algorithm for nonlinear equality constrained optimization*, in: P.T. Boggs, R.H. Byrd, and R.B. Schnabel, eds., Numerical Optimization (SIAM, Philadelphia, 1985), 71–82.

[8] T. Coleman and Y.Y. Li, *An interior point trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optimization **6** (1996), 418–445.

[9] A. Conn, N. Gould, A. Sartenaer, and Ph. Toint, *On iterated-subspace methods for nonlinear optimization*, in: J. Adams and J.L. Nazareth, eds., Linear and Nonlinear Conjugate Gradient-Related Methods (1996), 50–79.

[10] Y. H. Dai and Y.X. Yuan, *Nonlinear Conjugate Gradient Methods* (in Chinese), Shanghai Science and Technology Publisher, Shanghai, 2000.

[11] M. Elad, B. Matalon, and M. Zibulevsky, *Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization*, Appl. Comput. Harmon. Anal. **23** (2007) 346–367

[12] R. Fletcher, *Second order correction for nondifferentiable optimization*, in: G.A. Watson, ed., Numerical Analysis, Springer-Verlag, Berlin, (1982), 85–115.

[13] P.E. Gill and M.W. Leonard, *Reduced-Hessian quasi-Newton methods for unconstrained optimization*, SIAM J. Optim. **1** (2001), 209–237.

[14] R. Glowinski and A. Marrocco, *Sur l'approximation, par éléements finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*, Laboria, 1975.

[15] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods*, Vol. 15 of Studies in Mathematics and its Applications, North-Holland Publishing Co., Amsterdam, 1983. Applications to the numerical solution of boundary value problems, Translated from the French by B. Hunt and D. C. Spicer.

[16] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator-splitting Methods in Nonlinear Mechanics*, Vol. 9 of SIAM Studies in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1989.

[17] N.I.M. Gould, D. Orban, and Ph.L. Toint, *Numerical methods for large-scale nonlinear optimization*, Acta Numerica (2005), 299–361.

[18] Grapiglia, G.N., Yuan, J.Y., and Yuan, Y.X., *A subspace version of the Powell-Yuan trust-region algorithm for equality constrained optimization*, J. Operations Research Society of China **1** (2013), 425– 451.

[19] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput. **23** (2001), 517–541.

[20] Lee, J.H., *A Subspace Algorithm for Nonlinear Equality Constrained Optimization*, Ph.D. thesis, ICMSEC, AMSS, Chinese Academy of Science, Beijing, 2009.

[21] D.C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Mathematical Programming **45** (1989), 503–528.

[22] X. Liu, X. Wang, Z.W. Wen, and Y.X. Yuan, *On the convergence of the self-consistent field iteration in Kohn-Sham density function theory*, accepted by SIAM Journal on Matrix Analysis and Applications (2014).

[23] X. Liu, Z.W. Wen and Y. Zhang, *Limited memory block Krylov subspace optimization for computing dominant singular value decompositions*, SIAM Journal on Scientific Computing **35** (2013), A1641-A1668

[24] Y.Y. Li and S. Osher, *Coordinate descent optimization for ℓ^1 minimization with application to compressed sensing; a greedy algorithm*, Inverse Probl. Imaging **3** (2009), 487–503

[25] Z. Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl. **72** (1992), 7–35

[26] D.Q. Mayne and E. Polak, *A superlinearly convergent algorithm for constrained optimization problems*, Math. Prog. Study **16** (1982), 45–61.

[27] Q. Ni and Y.X. Yuan, *A subspace limited memory quasi-Newton algorithm for large-scale nonlinear bound constrained optimization*, Mathematics of Computations **66** (1997), 1509–1520.

[28] M.J.D. Powell and Y.X. Yuan, *A trust region algorithm for equality constrained optimization*, Mathematical Programming **49** (1991), 189–211.

[29] Y. Saad, *Iterative Methods for Sparse Linear Systems: 2nd Ed.*, SIAM, Philedalphia, 2003.

[30] A. Sit, Z.J. Wu, and Y.X. Yuan, *A geometric buildup algorithm for the solution of the distance geometry problem using least-squares approximation*, Bulletin of Mathematical Biology **71** (2009), 1914–1933.

[31] J. Stoer and Y.X. Yuan, *A subspace study on conjugate gradient algorithms*, ZAMM Z. angew. Math. Mech. **75** (1995), 69–77.

[32] W.Y. Sun and Y.X. Yuan, *Optimization Theory and Methods: Nonlinear Programming*, Springer Verlag, New York, 2006.

[33] Z.H. Wang, Z.W. Wen, and Y.X. Yuan, *A subspace trust region method for large scale unconstrained optimization*, in: Y.X. Yuan eds. Numerical Linear Algebra and Optimization (Science Press, Beijing/NewYork, 2004), pp. 265–274.

[34] Z.H. Wang and Y.X. Yuan, *A subspace implementation of quasi-Newton trust region methods for unconstrained optimization*, Numerische Mathematik **104** (2006), 241–269.

[35] Z.W. Wen, D. Goldfarb, and K. Scheinberg, *Block coordinate descent methods for semidefinite programming*, in: Handbook on Semidefinite, Conic and Polynomial Optimization International Series in Operations Research & Management Science **166** (2012), 533–564.

[36] Z.W. Wen, A. Milzarek, M. Ulbrich, and H.C. Zhang, *Adaptive regularized self-consistent field iteration with exact Hessian for electronic structure calculation*, SIAM Journal on Scientific Computing **35** (2013), pp. A1299–A1324.

[37] Z.W. Wen, D. Goldfarb, and W. Yin, *Alternating direction augmented Lagrangian methods for semidefinite programming*, Mathematical Programming Computation **2** (2010), 203–230

[38] Z.W. Wen, W. Yin, D. Goldfarb, and Y. Zhang, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation*, SIAM Journal on Scientific Computing **32** (2010), 1832–1857.

[39] X.C. Tai and J.C. Xu, *Global and uniform convergence of subspace correction methods for some convex optimization problems*, Math. Comp. **71** (2002), 105–124.

[40] J. Yang and Y. Zhang, *Alternating direction algorithms for l_1-problems in compressive sensing*, SIAM journal on scientific computing **33** (2011), 250–278.

[41] Y.X. Yuan, *Subspace techniques for nonlinear optimization*, in: R. Jeltsch, D.Q. Li and I. H. Sloan, eds., Some Topics in Industrial and Applied Mathematics (Series in Contemporary Applied Mathematics CAM 8) (Higher Education Press. Beijing, 2007), pp. 206–218.

[42] ———, *Subspace methods for large scale nonlinear equations and nonlinear least squares*, Optimization and Engineering **10** (2009), 207–218.

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhong Guan Cun Donglu 55, Beijing 100190, China

E-mail: yyx@lsec.cc.ac.cn

16. Control Theory and Optimization

Recent results around the diameter of polyhedra

Friedrich Eisenbrand

Abstract. The diameter of a polyhedron P is the largest distance of a pair of vertices in the edge-graph of P. The question whether the diameter of a polyhedron can be bounded by a polynomial in the dimension and number of facets of P remains one of most important open problems in convex geometry. In the last three years, there was an accelerated interest in this famous open problem which has lead to many interesting results and techniques, also due to a celebrated breakthrough of Santos disproving the Hirsch conjecture. Here, I want to describe a subset of these recent results and describe some open problems.

Mathematics Subject Classification (2010). Primary 52B11; Secondary 52B55.

Keywords. Polyhedra, convex geometry, linear optimization, polynomial Hirsch conjecture.

1. Introduction

Linear programming is among the most important concepts in applied mathematics, theoretical computer science and optimization. The task is to maximize a linear objective function subject to linear constraints

$$\max\{c^T x \colon Ax \leq b\}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}^d$ are the *constraint matrix*, *right-hand-side* and *objective-function* vector respectively.

The classical *simplex method*, described by George Dantzig [14], see also [11, 13, 37, 40] is one of the most effective methods to solve linear programming problems in practice. The algorithm is readily described but we need some terminology. A set of the form $P = \{x \in \mathbb{R}^d \colon Ax \leq b\}$ is called a *polyhedron*. Thus the set of *feasible solutions* of (1.1) is a polyhedron. A *vertex* of P is an element $x^* \in P$ such that there exist d linearly independent constraints of $Ax \leqslant b$ that are tight at x^*, i.e., satisfied by x^* with equality. Two vertices $x^* \neq y^*$ are *neighbors* of each other, if there exist $d-1$ linearly independent constraints that are tight at both.

Let us now assume that the polyhedron P of feasible solutions has vertices and that the linear program (1.1) is bounded. The simplex algorithm then starts at a vertex of P and moves to a neighbor with strictly larger objective function value. If no such neighbor exists, then the current vertex is an optimal solution of (1.1).

Linear programming can be solved in polynomial time with the *ellipsoid method* [27] or *interior-point* methods [26], see also [21]. The running time bound of these algorithms is however polynomial in n, d and the largest *binary encoding length* of a coefficient of

Proceedings of the International Congress of Mathematicians, Seoul, 2014

the input. The encoding length of the intermediate rational numbers in the course of the algorithm also remains polynomial.

There are computational problems where a dependence on the binary encoding length seems natural. The *Euclidean algorithm*, for example, requires a linear number of arithmetic operations to compute the *greatest common divisor* of two integers. In fact, Mansour et al. [32] show that a dependence on the binary encoding length is necessary in the computation-tree-model. Is such a dependence of the running time on the binary encoding length also necessary for linear programming? This is one of the most visible open problems in computer science and optimization, see also [42]. In complexity jargon, the question is whether there exists a *strongly polynomial-time* algorithm for linear programming. This is an algorithm that solves a linear program in time polynomial in n and d. Here the complexity of basic arithmetic operations does not count in the analysis. Yet the intermediate numbers should have binary encoding length that is polynomial in the encoding length of the input, see [21].

A vertex x^* might have several neighbors with better objective function value and there are *pivoting rules* that determine which improving neighbor to chose. There are deterministic, as well as randomized pivoting rules. For many of these pivoting rules, researchers were able to derive superpolynomial lower bounds on the (expected) number of iterations of the simplex algorithm [3, 20, 28, 34]. Still, more than 65 years after its publication, the simplex method continues to be a candidate for a strongly polynomial time algorithm for linear programming.

The simplex algorithm performs a walk on the *polyhedral graph* $G = (V, E)$ of P. The set of vertices V of G is the set of vertices of the polyhedron P and two vertices are joined by an edge $e = uv \in E$ if u and v are neighbors of P. This graph is connected. The *diameter* of P is the smallest natural number that bounds the length of a shortest path between any pair of vertices in this graph. The simplex algorithm that we described above can only terminate in polynomial time in d and n if this diameter can be bounded by a polynomial in d and n. This gives rise to the *diameter problem* which is one of the most prominent mysteries in convex and discrete geometry. It is in the focus of our interest.

> Can the diameter of a polyhedron $P = \{x \in \mathbb{R}^d \colon Ax \leqslant b\}$ be bounded by a polynomial in d and n?

The belief in a positive answer to this question is called the *polynomial Hirsch conjecture*. In the following, we denote the smallest upper bound on the diameter of a polyhedron in dimension d described by n inequalities by $\Delta(d, n)$.

The classical *Hirsch conjecture* itself, stated in 1957, was proposing the explicit linear bound

$$\Delta(d, n) \leq n - d. \tag{1.2}$$

This conjecture was disproved for unbounded polyhedra by Klee and Walkup [29] and it remained a very highly visible conjecture for polytopes (bounded polyhedra) since then. In a celebrated paper by Santos [38] it was recently disproved. As of today, the known counter examples for the Hirsch conjecture [33, 38] violate the Hirsch bound only by a small constant factor.

While, when the dimension d is fixed, the diameter can be bounded by a linear function of n [6, 30], for the general case the best upper bound is only quasipolynomial. Kalai and Kleitman [25] showed that $\Delta(d, n) \leq n^{1+\log d}$ holds. This was the best bound for more than 20 years. Recently, Todd [44] has provided an improvement of this bound. He shows that

$\Delta(d,n) \leq (n-d)^{\log d}$ by strengthening the arguments applied in the Kalai-Kleitman bound. Still, the gap between the linear lower bound and the quasi-polynomial upper bound is huge.

Research on the diameter problem has been very active recently and we cannot be exhaustive in our treatment of the subject. Instead we will focus on the following topics. The diameter problem was the subject of a recent polymath project [23]. Here, the idea was consider abstractions of polyhedral graphs for which the proof method of Kalai and Kleitman still applies. The project stipulated several nice results but also here, the gap between polynomial (quadratic in this case) and quasi-polynomial $n^{\log d+1}$ also still stands. There is an explicit conjecture of Hähnle that is tantalizing and we will describe it here. Also, we describe a recent result of Bonifas et al. [7]. Here, the authors have shown that the $\Delta(d,n)$ is bounded by $O\left(\mu^2 d^4 \log d\mu\right)$, where $A \in \mathbb{Z}^{n\times d}$ is an integral matrix whose absolute value of any sub-determinant is bounded by μ. This is a generalization of a result of Dyer and Frieze [17] for polyhedra that are described by a *totally unimodular* constraint matrix, i.e., $\mu = 1$. However, we describe this result in a more geometric setting proposed by Brunsch and Röglin [10] who were able to describe a strongly polynomial time algorithm to find a path joining two given vertices of a polyhedron if the constraint matrix satisfies a certain geometric property. We then describe a recent algorithmic result of Eisenbrand and Vempala [19] who found a randomized simplex algorithm to solve linear programs whose expected running time is polynomial in the parameter of Brunsch and Röglin and the dimension d.

2. Abstractions

Combinatorial abstractions of polyhedral graphs have been studied in the literature for a long time [1, 2, 24, 43]. Here we describe an abstraction that was presented by Eisenbrand et al. [18], see also [23]. Throughout we assume that the polyhedra we discuss have vertices.

A polyhedron $P \subseteq \mathbb{R}^d$ described by n inequalities whose diameter is largest among all those can be assumed to be *non-degenerate*. This means that each vertex of P satisfies exactly d of the inequalities defining P with equality. If P is degenerate, then one can perturb the right-hand-side vector b a bit such that the diameter of the new polyhedron is at least as large. This technique is for example described in [40].

The vertices of a non-degenerate polyhedron are thus uniquely described by d linearly independent inequalities that are tight at the vertex. Thus, from now on, we can identify each vertex by a d-element subset of $[n] = \{1,\ldots,n\}$.

The following is now crucial. Suppose $u \in \binom{[n]}{d}$ and $v \in \binom{[n]}{d}$ are two vertices. Then there exists a path in the polyhedral graph of P with endpoints u and v such that each intermediate vertex on this path contains $u \cap v$. This gives rise to the following abstraction from [18].

Definition 2.1. A *base abstraction* of *dimension* d with n facets is a graph $G = (V,E)$, where $V \subseteq \binom{[n]}{d}$ such that each pair of vertices $u, v \in V$ is connected by a path such that each intermediate vertex on that path contains $u \cap v$.

Notice that we do not specify the edges of the base abstraction. Instead one only requires the connectivity condition that is present in polyhedral graphs. This is *one single feature* that is common to the previously studied abstractions [1, 2, 24]. But already here, one can prove a Kalai-Kleitman bound.

Let $G = (V, E)$ be a base abstraction of dimension d with n facets and let u and v be two vertices that are furthest apart from each other. Let us consider a run of the breadth-first-search algorithm initiated with $S_1 = \{u\}$. Then breath first search finds the set $S_2 \subseteq V$, the set of vertices at distance one from u, $S_3 \subseteq V$ the set of vertices at distance 2 from u and so on, until it discovers v which is then in the set $S_{\ell+1}$, where ℓ is the distance of u and v.

Lemma 2.2. *Let x, y be vertices of the base abstraction where $x \in S_i$ and $y \in S_j$ with $i < j$, then each set S_k $i \leqslant k \leqslant j$ contains a vertex z with $z \supseteq x \cap y$.*

The following argument, proving Lemma 2.2 is from [18]. Since $G = (V, E)$ is a base abstraction, there exists a path with endpoints x and y such that each vertex on that path contains $x \cap y$. Following this path, the distance labels (distances from u) cannot jump up or down by more than 1. Thus each distance label between i and j must be observed. This implies the lemma.

Such collections of sets $S_1, \ldots, S_{\ell+1}$ are called *connected layer families*. A formal definition is as follows.

Definition 2.3. A d-dimensional connected layer family with n symbols is a collection of sets $S_1, \ldots, S_{\ell+1}$ such that

1. each S_i is a set of d-element subsets of $[n]$, i.e., $S_i \subseteq \binom{[n]}{d}$,
2. the S_i are disjoint and
3. for each $1 \leqslant i < k < j \leqslant \ell + 1$ and each $x \in S_i$ and $y \in S_j$ there exists a $z \in S_k$ with $z \supseteq x \cap y$.

The *height* of the connected layer family is $\ell + 1$. Let $D(d, n)$ be the maximum diameter of a d-dimensional base abstraction and $h(d, n)$ be the maximum height of a d-dimensional connected layer family. It is easy to see that $h(d, n) - 1 = D(d, n)$. The Kalai-Kleitman bound is easy to prove for connected layer families. We follow the presentation given in [18].

Theorem 2.4 (Kalai & Kleitman[25]). *The maximum height $h(d, n)$ of a d-dimensional connected layer family with n symbols is bounded by $n^{1+\log d}$.*

Proof. Let $S_1, \ldots, S_\ell$ be a connected layer family. Let $i \geqslant 0$ be maximal such that the union of the d-sets in $S_1 \cup \cdots \cup S_i$ contains at most $n/2$ symbols. Likewise, let $j \leqslant \ell + 1$ be minimal such that the union of the d-sets in $S_\ell \cup \cdots \cup S_j$ contains at most $n/2$ symbols. There must be a symbol $s \in [n]$ that is contained in some d-set in each set $S_{j+1}, \ldots S_{j-1}$.

Now we observe that $S_1, \ldots, S_i$ and $S_j, \ldots, S_\ell$ are d-dimensional connected layer families with at most $\lfloor n/2 \rfloor$ symbols each. Also, $S_{i+1}, \ldots S_{j-1}$ is a connected layer family. This also remains to be the case if we delete each d-set from the family $S_{j+1}, \ldots S_{j-1}$ that does not contain the symbol s and then delete s from each remaining d-set. This shows that the following recursion holds

$$h(d, n) \leqslant 2 \cdot h(d, \lfloor n/2 \rfloor) + h(d - 1, n - 1). \tag{2.1}$$

The bound is then proved by induction on n. Note that $h(1, n) = n$ and $h(d, n) = 0$ if $d > n$. Suppose now that $d, n \geqslant 2$. Applying (2.1) repeatedly, one obtains

$$h(d, n) \quad \leqslant \quad 2 \cdot h(d, \lfloor n/2 \rfloor) + h(d - 1, n)$$

$$\leqslant \quad 2 \cdot \sum_{i=2}^{d} h(i, \lfloor n/2 \rfloor) + h(1, n).$$

By induction, this is bounded by

$$\begin{aligned} h(d,n) &\leqslant 2(d-1)(2d)^{\log n - 1} + n \\ &= (2d)^{\log n - 1} \left(2(d-1) + n/((2d)^{\log n - 1})\right) \\ &\leqslant (2d)^{\log n} \end{aligned}$$

In the last inequality one uses $d \geqslant 2$ and thus $(2d)^{\log n - 1} \geqslant n^2/4$. Since $n \geqslant 2$ one can conclude $n/((2d)^{\log n - 1}) \leqslant 4/n \leqslant 2$. □

As far as lower bounds are concerned, Eisenbrand et al. [18] show the following theorem.

Theorem 2.5 ([18]). *The diameter of a d-dimensional base abstraction with n symbols $D(n/4, n) = \Omega(n^2/\log n)$.*

Also the linear bound on the diameter for fixed d of Barnette [6] and Larman [30] holds for the base abstractions [18]. Todd [44] has recently improved the Kalai-Kleitman bound in the setting of polyhedra. He showed that the bound can be tightened to $(n-d)^{\log d}$.

2.1. The Hähnle conjecture. It turns out that the base-abstraction can be further generalized such that the Kalai-Kleitman bound still holds. We describe this *generalized base abstraction* now. The vertices of the graph $G = (V, E)$ are now a subset of the degree-d monomials in $\mathbb{Z}[x_1, \ldots, x_n]$. For any $x^\alpha, x^\beta \in V$ we require that there exists a path such that $\gcd(x^\alpha, x^\beta)$ divides each vertex on that path. If no monomial is divisible by some x_i^2, then we are in the setting of our previous base-abstraction. The following examples have been discussed in [23].

Example 2.6. The set of vertices is the set of degree-d monomials that involve only two and consecutive variables. In other words

$$V = \{x_i^k x_{i+1}^{d-k} \colon 1 \leqslant i \leqslant n-1,\ 0 \leqslant k \leqslant d\}.$$

Two monomials x^α, x^β form an edge, if and only if $x^\alpha/x^\beta = x_i/x_{i+1}$ for some $i \in 1, \ldots, n-1$. In other words, the graph is a path of the form

$$x_1^d, x_1^{d-1}x_2, x_1^{d-2}x_2^2, \ldots, x_{n-1}^1 x_n^{d-1}, x_n^d.$$

The diameter of this graph is $n \cdot (d-1)$.

Example 2.7. The set of vertices is complete, i.e., comprises all monomials of degree d. We group these vertices into sets V_i for $i = d, \ldots, n \cdot d$ where V_i consists of all monomials x^α with $i = \prod_{j=1}^n j \cdot \alpha_j$. Each group V_i is a clique and we totally connect each pair of groups V_i, V_{i+1}, for $i = d, \ldots n \cdot d - 1$. Clearly, this satisfies the connectivity condition. The diameter of this graph is $(n-1) \cdot d$.

So far, no example of a generalized base abstraction is known, whose diameter is larger than $(n-1) \cdot d$. Nicolai Hähnle [23] proposed the following conjecture.

Conjecture 2.8 (Hähnle conjecture). *The diameter of a generalized base abstraction with n symbols in dimension d is bounded by $(n-1)\cdot d$.*

Santos [39] considered the following relaxation of the base abstraction. A *pure simplicial complex* of dimension $d-1$ is a family of d-subsets of $[n]$. Again, interpreting vertices of a polyhedron $P \subseteq \mathbb{R}^d$ with n facets via their defining inequalities, one obtains a *simplicial complex* of dimension $d-1$ on n vertices, where the vertices of the complex are the labels $\{1,\ldots,n\}$ of the facets of the polyhedron.

Two d-sets of the complex are adjacent if their intersection has $d-1$ elements. The connectivity condition of the base abstraction can be cast as follows. If one fixes a set $u \subseteq [n]$, then the sub-complex consisting of all d-sets containing u is connected.

Santos [39] shows that the diameter of pure simplicial complexes can be exponential in n and d. This proves that the connectivity condition is essential to derive the aforementioned upper bounds.

3. Diameter bounds for special cases

We now turn to the diameter problem for certain classes of polytopes. *Combinatorial optimization* problems can often be modeled as a linear programming problem over the convex hull of the characteristic vectors of the solutions. These characteristic vectors are vectors with components in $\{0,1\}$. A polytope that is the convex hull of $0/1$-vectors is called a $0/1$ *polytope*. Naddef [35] proved that the Hirsch conjecture holds true for $0/1$-polytopes. Orlin [36] provided a quadratic upper bound for flow-polytopes. Brightwell et al. [9] showed that the diameter of the transportation polytope is linear in n and d, and a similar result holds for the dual of a transportation polytope [5] and the axial 3-way transportation polytope [15].

The results on flow polytopes and classical transportation polytopes concern polyhedra defined by *totally unimodular matrices*, i.e., integer matrices whose sub-determinants are $0,\pm 1$. For such $P = \{x \in \mathbb{R}^d \colon Ax \leqslant b\}$ with A totally unimodular, Dyer and Frieze [17] have shown that the diameter is bounded by a polynomial in d and n. Interestingly, this bound holds *independent* of the right-hand-side vector $b \in \mathbb{R}^n$. The vector b can be irrational even. The bound is $O(n^{16}d^3(\log nd)^3)$. Their result is also algorithmic: they show that there exists a randomized simplex-algorithm that solves linear programs defined by totally unimodular matrices in polynomial time.

In [7] the authors improve and generalize the aforementioned bound of Dyer and Frieze. The authors show that the diameter of a polyhedron $P = \{x \in \mathbb{R}^d \colon Ax \leqslant b\}$, with $A \in \mathbb{Z}^{n\times d}$ is bounded by $O\left(\Delta^2 d^4 \log d\Delta\right)$. Here, Δ denotes the largest absolute value of a *sub-determinant* of A. If P is bounded, i.e., a *polytope*, then they show that the diameter of P is at most $O\left(\Delta^2 n^{3.5}\log n\Delta\right)$. Notice that this bound is independent of n, i.e., the number of rows of A. On the other hand, if A is totally unimodular, then $n = O(d^2)$ [22]. If the sub-determinants of an integer matrix are bounded by Δ, then $n = O(d^t)$ with $t = \Omega(\Delta)$ [4, 31].

3.1. A geometric generalization. Recently, Brunsch and Röglin [10] suggested the following generalization of integer matrices with bounded sub-determinant. Consider again a polyhedron

$$P = \{x \in \mathbb{R}^d \colon Ax \leqslant b\},$$

where $A \in \mathbb{R}^{n\times d}$ is of full-column-rank and has the following property.

The sign of the angle of a row of A to a subspace of $\mathbb{R}^d$ that is generated by $d-1$ other rows of A is at least δ.

How large is δ in terms of d and the largest sub-determinant? Let $A \in \mathbb{Z}^{n\times d}$ and suppose that all sub-determinants are bounded by Δ and let $a_1, \ldots, a_d$ be d linearly independent rows of A. The *adjoint* matrix $\widetilde{C} = (b_1, \ldots, b_d)$ of the matrix with rows $a_1, \ldots, a_d$ is an integer matrix with all components in $\{-\Delta, \ldots, \Delta\}$. The vector b_1 is orthogonal to $a_2, \ldots, a_d$ and $|a_1^T b_1| \geqslant 1$, since the vectors are integral. The distance of a_1 to the sub-space generated by $a_2, \ldots, a_d$ is thus at least $1/\|b_1\| \geqslant 1/(\Delta\sqrt{d})$ which means that $\delta \geqslant 1/(\Delta^2 d)$.

Brunsch and Röglin [10] have shown how to compute a path between two given vertices of P in time polynomial in n, d and $1/\delta$. The expected length of their path is bounded by $O(nd^2/\delta^2)$. Their algorithm is as follows.

Suppose that the rows of A are scaled in such a way that they are of unit length and suppose the two given vertices of P are u and v. In a first step, one computes index sets $B_u, B_v \subseteq \{1, \ldots, d\}$ such that the sub-matrices A_{B_u} and A_{B_v} are linearly independent and the corresponding inequalities $A_{B_u}x \leqslant b_{B_u}$ and $A_{B_v}x \leqslant b_{B_v}$ are satisfied with equality by u and v respectively. Now let $\lambda, \mu \in [0,1]^d$ be chosen independently and uniformly at random. This yields two vectors $c_u^T = \lambda^T A_u$ and $c^T v = \mu^T A_v$.

Brunsch and Röglin use the *shadow-vertex* pivoting rule [8] to walk from u to v along the edges of P. Consider the image of P under the *projection* $p(x) = (c_u^T x, c_v^T x)$. Walking upwards from $p(u)$ to $p(v)$ in this projection along the edges of this convex polygon corresponds to a walk on the polyhedral graph of P from u to v. Brunsch and Röglin [10] show that the expected number of different slopes of this projection is bounded by $O(nd^2/\delta^2)$. They also argue that w.h.p. no two edges have the same slope in this projection.

3.2. Bounding the diameter in terms of δ and d. In the following, we prove the bound of Bonifas et al. in the setting of Brunsch and Röglin, thereby obtaining a bound on the diameter that is polynomial in d and $1/\delta$. Let u and v be two vertices of P. We estimate the maximum number of iterations of two breadth-first-search explorations of the polyhedral graph, one initiated at u, the other initiated at v, until a common vertex is discovered. The diameter of P is at most twice this number of iterations. The main idea in the analysis is to reason about the normal cones of vertices of P and to exploit a certain volume expansion property.

Again, we can assume that $P = \{x \in \mathbb{R}^d \colon Ax \leqslant b\}$ is *non-degenerate*, i.e., each vertex has exactly d tight inequalities. Let $v \in V$ now be a vertex of P. The *normal cone* C_v of v is the set of all vectors $c \in \mathbb{R}^d$ such that v is an optimal solution of the linear program $\max\{c^T x \colon x \in \mathbb{R}^d,\ Ax \leqslant b\}$. The normal cone C_v of a vertex of v is a full-dimensional simplicial polyhedral cone. Two vertices v and v' are adjacent if and only if C_v and $C_{v'}$ share a facet. No two distinct normal cones share an interior point. Furthermore, if P is a polytope, then the union of the normal cones of vertices of P is the complete space $\mathbb{R}^d$.

We now define the *volume* of a set $U \subseteq V$ of vertices as the volume of the union of the normal cones of U intersected with the *unit ball* $B_d = \{x \in \mathbb{R}^d \colon \|x\|_2 \leqslant 1\}$, i.e.,

$$\operatorname{vol}(U) := \operatorname{vol}\left(\bigcup_{v\in U} C_v \cap B_d\right).$$

From now on, we will denote the normal cone of a vertex v intersected with B_d by C_v itself and call it *cone of* v.

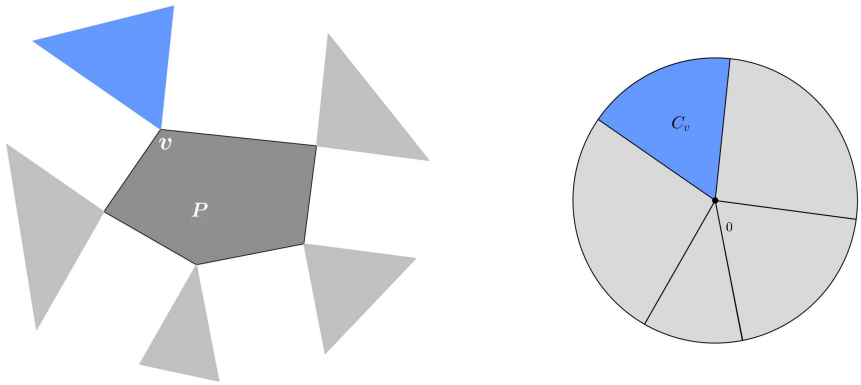


Figure 3.1. A polytope P, a vertex v of P and the cone C_v.

Consider an iteration of breadth-first-search. Let $I \subseteq V$ be the set of vertices that have been discovered so far. Breadth-first-search will next discover the neighborhood of I, which we denote by $\mathcal{N}(I)$. The key observation is that the volume is rapidly expanding. This is captured in the next lemma.

Lemma 3.1 (Bonifas et al. [7]). *Let $P = \{x \in \mathbb{R}^d \colon Ax \leqslant b\}$ be a polytope and let $I \subseteq V$ be a set of vertices with $\mathrm{vol}(I) \leqslant (1/2) \cdot \mathrm{vol}(B_d)$. Then the volume of the neighborhood of I is at least*

$$\mathrm{vol}(\mathcal{N}(I)) \geqslant \sqrt{\frac{2}{\pi}}(\delta/d^{1.5}) \cdot \mathrm{vol}(I).$$

Proof. Consider the vertices of I and the union of their cones

$$\bigcup_{v \in I} C_v.$$

This set has *exposed* facets, that are not covered by other cones of vertices. In the next iteration of breadth-first-search all these exposed facets must be covered by cones of neighbors. Let $A(I)$ be the area of these exposed facets and let $A(v)$ be the area of the facets of the cone C_v. One has the relation

$$\sum_{v \in \mathcal{N}(I)} A(v) \geqslant A(I).$$

In order to show rapid expansion of volume, we thus must bound the volume the cone of one vertex by its area from below and we must bound the volume of a set of vertices by its area from above.

Lets begin with the volume of one vertex. Consider a facet F of C_v. Since the distance of the vertex opposite to F is at least δ, we have

$$\mathrm{vol}(C_v) \geqslant \int_0^1 \left(\frac{x}{\delta}\right)^{d-1} A(F)\, dx = \delta/d \cdot A(F),$$

where $A(F)$ denotes the area of the facet F. Thus $\mathrm{vol}(C_v) \geqslant \delta/d^2 \cdot A(C_v)$.

How large can $\mathrm{vol}(I)$ be compared to $A(I)$? By a classical isoperimetric inequality this volume is largest for the convex hull of a spherical cap and 0. The area-to-volume ratio of such a spherical cone is smallest for the half-ball (remember that we require the volume of I to be at most the volume of the half-ball) and thus at least $\sqrt{2 \cdot n/\pi}$.

Thus

$$\sum_{v \in \mathcal{N}(I)} \mathrm{vol}(C_v) \geqslant \sum_{v \in \mathcal{N}(I)} \delta/d^2 \cdot A(C_v) \geqslant \delta/d^2 A(I) \geqslant \delta/d^2 \sqrt{\frac{2 \cdot d}{\pi}} \mathrm{vol}(I),$$

and the bound follows.

□

With this volume expansion lemma, we can prove a bound on the diameter of P that is polynomial in $1/\delta$ and d.

Theorem 3.2 (Bonifas et al. [7]). *Let $P = \{x \in \mathbb{R}^d \colon Ax \leqslant b\}$ be a polytope where the sign of the angle of any row of A to the subspace generated by $d-1$ other rows of A is at least δ. The diameter of P is bounded by $O\left(d^{2.5}/\delta \cdot \ln(d/\delta)\right)$.*

Proof. We begin breadth-first-search with the vertex u and estimate the number of steps until the volume of visited vertices exceeds $1/2 \cdot \mathrm{vol}(B_d)$. To this end, let $I_0 = \{u\}$ and let I_j be the set of vertices that have been discovered in the first j iterations of breadth-first-search.

From Lemma 3.1 one has

$$\mathrm{vol}(I_j) \geqslant \left(1 + \sqrt{\frac{2}{\pi}}(\delta/d^{1.5})\right)^j \cdot \mathrm{vol}(C_v).$$

The volume of I_j cannot exceed the volume of the ± 1 cube which is 2^d. Also the volume of the cone C_v is at least $\mathrm{vol}(C_v) \geqslant \delta^d/d!$. This is, because the determinant of the d rows of A that generate C_v is at least δ^d and the simplex generated by these rows and 0 is contained in C_v. Thus, one has

$$(2 \cdot d/\delta)^d \geqslant \left(1 + \sqrt{\frac{2}{\pi}}(\delta/d^{1.5})\right)^j$$

and thus

$$d \cdot \ln(2 \cdot d/\delta) \geqslant j \cdot \ln\left(1 + \sqrt{\frac{2}{\pi}}(\delta/d^{1.5})\right).$$

For $0 \leqslant \xi \leqslant 1$ one has $\ln(1+\xi) \geqslant \xi/2$ and thus the inequality above implies

$$d \cdot \ln(2 \cdot d/\delta) \geqslant j \cdot \left(\sqrt{\frac{1}{2 \cdot \pi}}(\delta/d^{1.5})\right),$$

from which the bound follows.

□

3.3. A simplex algorithm that is polynomial in d and $1/\delta$. In a recent paper [19], the authors have described a simplex algorithm to solve linear programs

$$\max\{c^T x\colon x \in \mathbb{R}^d, Ax \leqslant b\}$$

that has expected running time being polynomial in $1/\delta$ and d. The algorithm can be understood as an algorithmic realization of the diameter bound of Bonifas et al. [7] and extends a previous randomized dual-simplex algorithm of Dyer and Frieze [17] for the case of totally-unimodular constraint matrices.

Theorem 3.3 ([19]). *There is a variant of the simplex algorithm that solves a linear program with n constraints in $\mathbb{R}^d$ with probability at least $3/4$ using $\mathrm{poly}(d, 1/\delta)$ pivots. The pivot probabilities can be computed in time $\mathrm{poly}(d, \log(1/\delta))$.*

The algorithm performs a random walk on a sub-division of the cones of the polytope. If one starts at a cone of a given feasible solution, then the direct goal would be to walk to the cone that contains the objective function vector c. Here, we assume that c is a vector of ℓ_2-norm one. Instead, the algorithm stops already if it is in a cone that is sufficiently close to c. It finds a point c' in the unit sphere such that:

a) $\|c - c'\| < \delta/(2 \cdot n)$, and

b) we know a vertex v of P with $c' \in C_v$.

Once such a c' has been identified, one can identify at least one element of the optimal basis of the linear program by applying a sensitivity-result of Cook et al. [12]. This inequality can then be set to equality. The δ of the resulting $d-1$-dimensional linear program can only grow.

Consider the function $f\colon \mathbb{R}^d \to \mathbb{R}$ defined as $f(x) = e^{(c^T x)/t_0}$. For a cone C, let $f(C) = \int_{x \in C} f(x)\, dx$. The random walk described in [19], if run for a polynomial number of steps, is located in the cone C_v with probability proportional to its measure $f(C_v)/f(B_d)$. By choosing $t_0 = \delta^2/(16n^3)$ the cone in which the walk stops contains a point close to c satisfying a) and b) with high probability and one can identify an element of the optimal basis.

The sub-division of the original cones is carried out in such a way that the ratios of measures of neighboring cones is bounded from below by a polynomial in $1/d$ and δ. Then, again using an isoperimetric inequality [16] the *conductance* of the random walk is also bounded from below by a polynomial in δ and $1/d$. This implies [41] that the random walk is rapidly mixing.

We believe that it is an interesting open problem to find a deterministic simplex algorithm that runs in time that is polynomial in d and $1/\delta$.

References

[1] I. Adler, *Lower bounds for maximum diameters of polytopes*, Mathematical Programming Study, (1) (1974), 11–19. Pivoting and extensions.

[2] I. Adler, G. Dantzig, and K. Murty, *Existence of A-avoiding paths in abstract polytopes*, Mathematical Programming Study, (1) (1974), 41–42. Pivoting and extensions.

[3] N. Amenta and G. M. Ziegler, *Deformed products and maximal shadows of polytopes*, In Advances in discrete and computational geometry (South Hadley, MA, 1996), volume 223 of Contemp. Math., Amer. Math. Soc., Providence, RI, 1999, pp. 57–90.

[4] R. Anstee, *Forbidden configurations, discrepancy and determinants*, European Journal of Combinatorics, **11**(1) (1990), 15–19.

[5] M. L. Balinski, *The Hirsch conjecture for dual transportation polyhedra*, Math. Oper. Res., **9**(4) (1984), 629–633.

[6] D. Barnette, *An upper bound for the diameter of a polytope*, Discrete Math., **10** (1974), 9–13.

[7] N. Bonifas, M. Di Summa, F. Eisenbrand, N. Hähnle, and M. Niemeier, *On subdeterminants and the diameter of polyhedra*, In Proceedings of the 28th annual ACM symposium on Computational geometry, SoCG '12, 2012, pp. 357–362.

[8] K.-H. Borgwardt, *The average number of pivot steps required by the simplex-method is polynomial*, Zeitschrift für Operations Research. Serie A. Serie B, **26**(5) (1982), A157–A177.

[9] G. Brightwell, J. van den Heuvel, and L. Stougie, *A linear bound on the diameter of the transportation polytope*, Combinatorica, **26**(2) (2006), 133–139.

[10] T. Brunsch and H. Röglin, *Finding short paths on polytopes by the shadow vertex algorithm*, In Automata, Languages, and Programming, Springer, 2013, pp. 279–290.

[11] V. Chvátal, *Linear programming*. W. H. Freeman and Company, 1983.

[12] W. Cook, A. M. H. Gerards, A. Schrijver, and E. Tardos, *Sensitivity theorems in integer linear programming*, Mathematical Programming, **34** (1986), 251–264.

[13] G. B. Dantzig, *Linear programming and extensions*, Princeton university press, 1965.

[14] ———, *Maximization of a linear function of variables subject to linear inequalities*, In Koopmans, T. C., editor, Activity Analysis of Production and Allocation, John Wiley & Sons, New York, 1951, pp. 339–347.

[15] J. A. De Loera, E. D. Kim, S. Onn, and F. Santos, *Graphs of transportation polytopes*, J. Combin. Theory Ser. A, **116**(8) (2009), 1306–1325.

[16] M. Dyer and A. Frieze, *Computing the volume of convex bodies: A case where randomness provably helps*, Probabilistic combinatorics and its applications, **44** (1991), 123–70.

[17] ———, *Random walks, totally unimodular matrices, and a randomised dual simplex algorithm*, Mathematical Programming, **64**(1, Ser. A) (1994), 1–16.

[18] F. Eisenbrand, N. Hähnle, A. Razborov, and T. Rothvoß, *Diameter of polyhedra: Limits of abstraction*, Mathematics of Operations Research, **35**(4) (2010), 786–794.

[19] F. Eisenbrand and S. Vempala, *Geometric random edge*, arXiv preprint arXiv:1404.1568, (2014).

[20] O. Friedmann, T. D. Hansen, and U. Zwick, *Subexponential lower bounds for randomized pivoting rules for the simplex algorithm*, In STOC'11—Proceedings of the 43rd ACM Symposium on Theory of Computing, ACM, New York, 2011, pp. 283–292.

[21] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, volume 2 of Algorithms and Combinatorics, Springer, 1988.

[22] I. Heller et al, *On linear systems with integral valued solutions*, George Washington University, Logistics Research Project, 1956.

[23] G. Kalai, *Polymath III: The polynomial Hirsch conjecture, combinatorics and more*, see http://gilkalai.wordpress.com/.

[24] _____, *Upper bounds for the diameter and height of graphs of convex polyhedra*, Discrete Comput. Geom., **8**(4) (1992), 363–372.

[25] G. Kalai and D. J. Kleitman, *A quasi-polynomial bound for the diameter of graphs of polyhedra*, Bull. Amer. Math. Soc. (N.S.), **26**(2) (1992), 315–316.

[26] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica, **4**(4) (1984), 373–395.

[27] L. Khachiyan, *A polynomial algorithm in linear programming*, Doklady Akademii Nauk SSSR, **244** (1979), 1093–1097.

[28] V. Klee and G. J. Minty, *How good is the simplex algorithm?*, In Inequalities, III (Proc. Third Sympos., Univ. California, Los Angeles, Calif., 1969; dedicated to the memory of Theodore S. Motzkin), Academic Press, New York, 1972, pp. 159–175.

[29] V. Klee and D. W. Walkup, *The d-step conjecture for polyhedra of dimension $d < 6$*, Acta Math. 133, 1967, pp. 53–78.

[30] D. G. Larman, *Paths of polytopes*, Proc. London Math. Soc. (3), **20** (1970), 161–178.

[31] J. Lee, *The incidence structure of subspaces with well-scaled frames*, Journal of Combinatorial Theory, Series B, **50**(2) (1990), 265–287.

[32] Y. Mansour, B. Schieber, and P. Tiwari, *A lower bound for integer greatest common divisor computations*, Journal of the ACM (JACM), **38**(2) (1991), 453–471.

[33] B. Matschke, F. Santos, and C. Weibel, *The width of 5-dimensional prismatoids*, arXiv preprint arXiv:1202.4701, 2012.

[34] K. G. Murty, *Computational complexity of parametric linear programming*, Mathematical Programming, **19**(2) (1980), 213–219.

[35] D. Naddef, *The Hirsch conjecture is true for (0,1)-polytopes*, Mathematical Programming, **45** (1989), 109–110.

[36] J. B. Orlin, *A polynomial time primal network simplex algorithm for minimum cost flows*, Mathematical Programming, **78**(2, Ser. B) (1997), 109–129. Network optimization: algorithms and applications (San Miniato, 1993).

[37] M. Padberg, *Linear optimization and extensions*, volume 12 of Algorithms and Combinatorics. Springer, 1995.

[38] F. Santos, *A counterexample to the Hirsch conjecture*, Ann. of Math. (2), **176**(1) (2012), 383–412.

[39] _____, *Recent progress on the combinatorial diameter of polytopes and simplicial complexes*, Top, **21**(3) (2013), 426–460.

[40] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley, 1986.

[41] A. Sinclair and M. Jerrum, *Approximate counting, uniform generation and rapidly mixing Markov chains*, Information and Computation, **82**(1) (1989), 93–133.

[42] S. Smale, *Mathematical problems for the next century*, The Mathematical Intelligencer, **20**(2) (1998), 7–15.

[43] M. J. Todd, *A generalized complementary pivoting algorithm*, Mathematical Programming, **6**(1) (1974), 243–263.

[44] _____, *An improved Kalai-Kleitman bound for the diameter of a polyhedron*, arXiv preprint arXiv:1402.3579, 2014.

EPFL, Station 8 , 1015 Lausanne, Switzerland
E-mail: friedrich.eisenbrand@epfl.ch

Optimization over polynomials: Selected topics

Monique Laurent

Abstract. Minimizing a polynomial function over a region defined by polynomial inequalities models broad classes of hard problems from combinatorics, geometry and optimization. New algorithmic approaches have emerged recently for computing the global minimum, by combining tools from real algebra (sums of squares of polynomials) and functional analysis (moments of measures) with semidefinite optimization. Sums of squares are used to certify positive polynomials, combining an old idea of Hilbert with the recent algorithmic insight that they can be checked efficiently with semidefinite optimization. The dual approach revisits the classical moment problem and leads to algorithmic methods for checking optimality of semidefinite relaxations and extracting global minimizers. We review some selected features of this general methodology, illustrate how it applies to some combinatorial graph problems, and discuss links with other relaxation methods.

Mathematics Subject Classification (2010). Primary 44A60, 90C22, 90C27, 90C30; Secondary 14P10, 13J30, 15A99.

Keywords. Positive polynomial, sum of squares, moment problem, combinatorial optimization, semidefinite optimization, polynomial optimization.

1. Introduction

Polynomial optimization. We consider the following *polynomial optimization problem*: given multivariate polynomials $f, g_1, \ldots, g_m \in \mathbb{R}[\mathbf{x}_1, \ldots, \mathbf{x}_n]$, compute the infimum of the polynomial function f over the basic closed semialgebraic set

$$K = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \ldots, g_m(x) \geq 0\} \tag{1.1}$$

defined by the polynomial inequalities $g_j(x) \geq 0$. That is, compute

$$f_{\min} := \inf_{x \in K} f(x) = \inf\{f(x) : g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}. \tag{P}$$

This is a in general hard, nonlinear and nonconvex optimization problem which models a multitude of problems from combinatorics, geometry, control and many other areas of mathematics and its applications.

Well established methods from nonlinear optimization can be used to tackle problem (P), which however can only guarantee to find *local* minimizers. Exploiting the fact that the functions f, g_j are polynomials, new algorithmic methods have emerged in the past decade that may permit to find *global* minimizers. These methods rely on using algebraic tools

Proceedings of the International Congress of Mathematicians, Seoul, 2014

(*sums of squares of polynomials*) and analytic tools (*moments of measures*) combined with *semidefinite optimization*.

In a nutshell, sums of squares of polynomials are used to certify positive polynomials, the starting point being that finding $f_{\min}$ amounts to finding the largest scalar λ for which the polynomial $f - \lambda$ is nonnegative on the set K. The key insight is that, while it is hard to test whether a polynomial f is nonnegative, one can test whether f can be written as a sum of squares of polynomials using semidefinite optimization.

Moments of measures are used to model the nonlinearities arising in polynomial functions, the starting point being that finding $f_{\min}$ amounts to finding a positive measure μ on the set K minimizing the integral $\int_K f(x)d\mu = \sum_\alpha f_\alpha \int_K x^\alpha d\mu$. These moments are used to build certain positive semidefinite Hankel type matrices. The key feature of these matrices is that they permit to certify optimality and to find the global minimizers of problem (P) (under certain conditions).

Semidefinite optimization is a wide generalization of the classical tool of linear optimization, where vector variables are replaced by matrix variables constrained to be positive semidefinite. In other words semidefinite optimization is linear optimization over affine sections of the cone of positive semidefinite matrices. The crucial property is that there are efficient algorithms for solving semidefinite programs (to any arbitrary precision).

Sums of squares and moment based methods permit to construct convex relaxations for the original problem (P), whose optimal values can be computed with semidefinite optimization and provide hiearchies of bounds for the global minimum $f_{\min}$. Convergence properties rely on real algebraic results (giving sums of squares certificates for positive polynomials), and optimality conditions and techniques for extracting global minimizers rely on functional analytic results for moment sequences combined with commutative algebra. Hence these methods have their roots in some classical mathematical results, going back to work of Hilbert about positive polynomials and sums of squares and to work on the classical moment problem in the early 1900's. They also use some recent algebraic and functional analytic developments combined with some modern optimization techniques that emerged since a few decades.

Some combinatorial examples. When all polynomials in (P) are linear, problem (P) boils down to linear programming:

$$\min\{c^{\mathsf{T}}x : Ax \geq b\}, \tag{LP}$$

well known to be solvable in polynomial time. However, when adding in (LP) the quadratic conditions $x_i^2 = x_i$ on the variables, we get 0/1 integer linear programming (ILP), which is hard. Instances of polynomial optimization problems arise naturally from combinatorial problems.

Consider for instance the *partition problem*, which asks whether a given sequence $a_1, \ldots, a_n$ of integers can be partitioned into two classes with equal sums, well known to be NP-complete [31]. This amounts to deciding whether the minimum over $\mathbb{R}^n$ of the polynomial $f = (\sum_{i=1}^n a_i \mathbf{x}_i)^2 + \sum_{i=1}^n (\mathbf{x}_i^2 - 1)^2$ is equal to 0.

We now mention other NP-hard problems, dealing with cuts, stable sets, graph colorings, and matrix copositivity, to which we will come back later in the paper.

Max-cut. Consider a graph $G = (V, E)$ with edge weights $w = (w_{ij}) \in \mathbb{R}^E$. The *max-cut problem* asks for a partition of the vertices of G into two classes in such a way that the total weight of the edges crossing the partition is maximum. Encoding partitions by vectors in

$\{\pm 1\}^V$, we obtain the following polynomial optimization problem:

$$\mathrm{mc}(G,w) = \max_{x\in\mathbb{R}^V}\{\sum_{\{i,j\}\in E}(w_{ij}/2)(1-x_ix_j) : x_i^2 = 1\ (i\in V)\}, \tag{1.2}$$

which models the max-cut problem. A basic idea to arrive at a semidefinite relaxation of problem (1.2) is to observe that, for any $x \in \{\pm 1\}^V$, the matrix $X = xx^{\mathsf{T}}$ is positive semidefinite and all its diagonal entries are equal to 1. This leads to the following problem:

$$\mathrm{sdp}(G,w) = \max_{X\in\mathbb{R}^{V\times V}}\{\sum_{\{i,j\}\in E}(w_{ij}/2)(1-X_{ij}) : X_{ii} = 1\ (i\in V),\ X\succeq 0\}, \tag{1.3}$$

where the notation $X \succeq 0$ means that X is symmetric positive semidefinite (i.e., $x^{\mathsf{T}}Xx \geq 0$ for all $x \in \mathbb{R}^V$). Of course if we would add the condition that X must have rank 1, then this would be a reformulation of the max-cut problem, thus intractable. The program (1.3) is an instance of semidefinite program and it can be solved in polynomial time (to any precision) as will be recalled below. This is the semidefinite program used by Goemans and Williamson [34] in their celebrated 0.878-approximation algorithm for max-cut. They show that for nonnegative edge weights the integrality gap $\mathrm{mc}(G,w)/\mathrm{sdp}(G,w)$ is at least 0.878 and they introduce a novel rounding technique to produce a good cut from an optimal solution to the semidefinite program (1.3). This is a breakthrough application of semidefinite optimization to the design of approximation algorithms, which started much of the research activity in this field (see e.g. [32]).

Stable sets and colorings. A stable set in a graph $G = (V, E)$ is a set of vertices that does not contain any edge. The *stability number* $\alpha(G)$ of G is the maximum cardinality of a stable set in G. It can be computed with any of the following two programs:

$$\alpha(G) = \max_{x\in\mathbb{R}^V}\sum_{i\in V}x_i \ \text{ s.t. } x_ix_j = 0\ (\{i,j\}\in E),\ x_i^2 = x_i\ (i\in V), \tag{1.4}$$

$$\frac{1}{\alpha(G)} = \min_{x\in\mathbb{R}^V}\ x^T(I+A_G)x \ \text{ s.t. } \sum_{i\in V}x_i = 1,\ x_i\geq 0\ (i\in V), \tag{1.5}$$

where A_G is the adjacency matrix of G (see [24] for (1.5)). As computing $\alpha(G)$ is NP-hard, we find again that problem (P) is hard as soon as some nonlinearities occur, either in the constraints (as in (1.4)), or in the objective function (as in (1.5)). Both formulations are useful to construct hierarchies of bounds for $\alpha(G)$.

The *chromatic number* $\chi(G)$ of G is the minimum number of colors needed to color the vertices so that adjacent vertices receive distinct colors. There is a classic reduction to the stability number. Consider the cartesian product $G\Box K_k$ of G and the complete graph on k nodes, whose edges are the pairs $\{(i,h),(j,h')\}$ with $i = j \in V$ and $h \neq h' \in [k]$, or with $\{i,j\} \in E$ and $h = h' \in [k]$. Then a stable set in the cartesian product $G\Box K_k$ corresponds to a subset of vertices of G that can be properly colored with k colors. Hence k colors suffice to properly color all the vertices of G precisely when $\alpha(G\Box K_k) = |V|$. Therefore, $\chi(G)$ is the smallest integer k for which $\alpha(G\Box K_k) = |V|$. This reduction will be useful for deriving hierarchies of bounds for $\chi(G)$ from bounds for $\alpha(G)$.

A well known bound for both $\alpha(G)$ and $\chi(G)$ is provided by the celebrated *theta number*

$\vartheta(G)$ of Lovász [70], defined by the following semidefinite program:

$$\vartheta(G) = \max_{X \in \mathbb{R}^{V \times V}} \{ \sum_{i,j \in V} X_{ij} : \mathrm{Tr}(X) = 1,\ X_{ij} = 0\ (\{i,j\} \in E),\ X \succeq 0\}. \tag{1.6}$$

The following inequalities hold, known as *Lovász' sandwich inequalities*:

$$\alpha(G) \leq \vartheta(G) \leq \chi(\overline{G}) \ \text{ and } \ \omega(G) \leq \vartheta(\overline{G}) \leq \chi(G). \tag{1.7}$$

Here, $\overline{G}$ is the complement of G and $\omega(G) = \alpha(\overline{G})$ is the maximum cardinality of a clique (a set of pairwise adjacent vertices) in G. The inequality $\alpha(G) \leq \vartheta(G)$ is easy: any stable set S of G gives a feasible solution $X = \chi^S(\chi^S)^{\mathsf{T}}/|S|$ of the program (1.6), where $\chi^S \in \{0,1\}^V$ is the characteristic vector of S.

A graph G is called *perfect* if $\omega(H) = \chi(H)$ for every induced subgraph H of G. Chudnovsky et al. [14] showed that a graph G is perfect if and only if it does not contain an odd cycle of length at least five or its complement as an induced subgraph. In view of (1.7), we have $\alpha(G) = \vartheta(G)$ and $\chi(G) = \vartheta(\overline{G})$ for perfect graphs. Therefore, both parameters $\alpha(G)$ and $\chi(G)$ can be computed in polynomial time for perfect graphs, via the computation of the theta number, using semidefinite optimization. Moreover, maximum stable sets and minimum graph colorings can also be found in polynomial time [36]. This is an early breakthrough application of semidefinite optimization to combinatorial optimization and as of today no other efficient algorithm is known for these problems.

One can strengthen the theta number toward $\alpha(G)$ by adding in program (1.6) the nonnegativity constraint $X \geq 0$ on the entries of X (leading to the parameter $\vartheta'(G)$), and toward $\chi(G)$ by replacing the constraint $X_{ij} = 0$ by $X_{ij} \leq 0$ for all edges (leading to the parameter $\vartheta^+(G)$). Thus we have:

$$\alpha(G) \leq \vartheta'(G) \leq \vartheta(G) \leq \vartheta^+(G) \leq \chi(\overline{G}). \tag{1.8}$$

We will see how to build hierarchies of bounds toward $\alpha(G)$ and $\chi(G)$ strenghtening the parameters ϑ' and ϑ^+, using the sums of squares and moment approaches.

Copositive matrices. Another interesting instance of unconstrained polynomial optimization is *testing matrix copositivity*, which is a hard problem [27, 74]. Recall that a symmetric $n \times n$ matrix M is called *copositive* if the quadratic form $x^{\mathsf{T}}Mx$ is nonnegative over the nonnegative orthant $\mathbb{R}^n_+$ or, equivalently, the polynomial $f_M = \sum_{i,j=1}^n M_{ij}\mathbf{x}_i^2\mathbf{x}_j^2$ is nonnegative over $\mathbb{R}^n$. Starting with the formulation (1.5) of the stability number $\alpha(G)$, it follows that $\alpha(G)$ can also be computed with the following copositive program:

$$\alpha(G) = \min_{\lambda \in \mathbb{R}} \{\lambda : \lambda(I + A_G) - J \ \text{ is copositive}\}, \tag{1.9}$$

where J is the all-ones matrix. By using sums of squares certificates for certifying matrix copositivity, one can define a hierarchy of cones approximating the copositive cone, which can also be used to define hierarchies of semidefinite bounds for the parameters $\alpha(G)$ and $\chi(G)$.

This paper. The field of polynomial optimization has grown considerably in the recent years. It has roots in early work of Shor [97] and later of Nesterov [75], and the foundations were laid by the groundworks of Lasserre [53, 54] and Parrilo [82, 83]. The monograph of

Lasserre [56], our overview [68] and the handbook [1] can serve as a general source about polynomial optimization. We also refer to the monographs [72, 85] and to the overview [91] for an in-depth treatment of real algebraic aspects, and to the monograph [9] for links to convexity.

In this paper we will discuss only a small selection of results from this field. Inevitably we cannot make full references to the literature and we apologize for all omissions. We will treat some subjects where we have done some (modest) contributions and our choices are biased, in particular, toward properties of the moment relaxations and toward hierarchies of semidefinite bounds for combinatorial problems. Our interest in polynomial optimization was stirred by the work [54] of Lasserre explaining how his method applies to 0/1 linear programming and we are grateful to Jean Lasserre for his inspiring work. We realized that his approach has tight links with lift-and-project methods for combinatorial optimization. This in turn inspired us to show finite convergence for polynomial optimization over finite varieties, to give simple real algebraic proofs for several results about flat extensions of moment matrices, and to investigate hierarchies for combinatorial graph parameters.

The paper is organized as follows. We begin with preliminaries about semidefinite optimization and sums of squares of polynomials. Then we present the sums of squares and moment approaches for polynomial optimization, with a special focus on the properties of moment matrices that allow to certify optimality and extract global optimizers. Then some selected applications are discussed: for computing real roots of polynomial equations, for designing hierarchies of semidefinite approximations for the stability number and the chromatic number, and for approximating matrix copositivity, again with application to approximating graph parameters. We conclude with mentioning some other research directions where hierarchies of semidefinite relaxations are also being increasingly used.

2. Preliminaries

Notation. $\mathbb{N} = \{0, 1, 2, \ldots\}$ is the set of nonnegative integers, $\mathbb{N}^n_t$ consists of the sequences $\alpha \in \mathbb{N}^n$ with $|\alpha| := \sum_{i=1}^n \alpha_i \leq t$ for $t \in \mathbb{N}$ and, for $\alpha \in \mathbb{N}^n$, $\mathbf{x}^\alpha$ denotes the monomial $\mathbf{x}_1^{\alpha_1} \cdots \mathbf{x}_n^{\alpha_n}$ with degree $|\alpha|$. (We use boldface letters $\mathbf{x}, \mathbf{x}_i, ..$ to denote variables.) $\mathbb{R}[\mathbf{x}_1, \ldots, \mathbf{x}_n] = \mathbb{R}[\mathbf{x}]$ is the ring of polynomials in n variables and $\mathbb{R}[\mathbf{x}]_t$ its subspace of polynomials with degree $\leq t$. The vector $[\mathbf{x}]_t = (\mathbf{x}^\alpha)_{\alpha \in \mathbb{N}^n_t}$ lists the monomials of degree at most t (in some given order) and, for a polynomial $f \in \mathbb{R}[\mathbf{x}]_t$, the vector $\mathbf{f} = (f_\alpha)_{\alpha \in \mathbb{N}^n_t}$ lists the coefficients of f (in the same order), so that $f = \sum_\alpha f_\alpha \mathbf{x}^\alpha = \mathbf{f}^\mathsf{T}[\mathbf{x}]_t$.

Given polynomials $g_1, \ldots, g_m$, we let $\mathcal{I} = (g_1, \ldots, g_m)$ denote the ideal that they generate and, for an integer t, $\mathcal{I}_t$ denotes its truncation at degree t, which consists of all polynomials $\sum_{j=1}^m p_j g_j$ with $p_j \in \mathbb{R}[\mathbf{x}]$ and $\deg(p_j g_j) \leq t$.

A polynomial f is a *sum of squares* (*sos*) if $f = g_1^2 + \ldots + g_m^2$ for some polynomials $g_1, \ldots, g_m$. $\Sigma[\mathbf{x}]$ contains all sums of squares of polynomials and we set $\Sigma[\mathbf{x}]_t = \Sigma[\mathbf{x}] \cap \mathbb{R}[\mathbf{x}]_t$. $\mathcal{P}(K)$ contains all polynomials f that are nonnegative over a given set $K \subseteq \mathbb{R}^n$, i.e., $f(x) \geq 0$ for all $x \in K$, also abbreviated as $f \geq 0$ on K.

Ideals and varieties. Consider an ideal $\mathcal{I} \subseteq \mathbb{R}[\mathbf{x}]$. The sets

$$\sqrt{\mathcal{I}} := \{f \in \mathbb{R}[\mathbf{x}] \mid f^k \in \mathcal{I} \text{ for some integer } k \geq 1\},$$

$$\sqrt[\mathbb{R}]{\mathcal{I}} := \{f \in \mathbb{R}[\mathbf{x}] \mid f^{2k} + p_1^2 + \ldots + p_m^2 \in \mathcal{I} \text{ for some } k \geq 1,\ p_1, \ldots, p_m \in \mathbb{R}[\mathbf{x}]\}$$

are called, respectively, the *radical* and the *real radical* of $\mathcal{I}$. Moreover, the sets

$$V(\mathcal{I}) = \{x \in \mathbb{C}^n : f(x) = 0 \; \forall f \in \mathcal{I}\}, \quad V_{\mathbb{R}}(\mathcal{I}) = V(\mathcal{I}) \cap \mathbb{R}^n$$

are, respectively, the *(complex) variety* and the *real variety* of the ideal $\mathcal{I}$. If $\mathcal{I} = (g_1, \ldots, g_m)$ is the ideal generated by a set of polynomials $g_1, \ldots, g_m$, then $V(\mathcal{I})$ consists of all their common complex roots while $V_{\mathbb{R}}(\mathcal{I})$ consists of their common real roots. The *vanishing ideal* of a set $V \subseteq \mathbb{C}^n$ is the set of polynomials

$$\mathcal{I}(V) = \{f \in \mathbb{R}[\mathbf{x}] : f(x) = 0 \; \forall x \in V\}.$$

The sets $\mathcal{I}(V)$, $\sqrt{\mathcal{I}}$ and $\sqrt[\mathbb{R}]{\mathcal{I}}$ are all ideals in $\mathbb{R}[\mathbf{x}]$ and they satisfy the inclusions:

$$\mathcal{I} \subseteq \sqrt{\mathcal{I}} \subseteq \mathcal{I}(V(\mathcal{I})) \text{ and } \mathcal{I} \subseteq \sqrt[\mathbb{R}]{\mathcal{I}} \subseteq \mathcal{I}(V_{\mathbb{R}}(\mathcal{I})).$$

The ideal $\mathcal{I}$ is called *radical* if $\mathcal{I} = \sqrt{\mathcal{I}}$ and *real radical* if $\mathcal{I} = \sqrt[\mathbb{R}]{\mathcal{I}}$. For instance, the ideal $\mathcal{I} = (\mathbf{x}^2)$ is not radical since $\mathbf{x} \in \sqrt{\mathcal{I}} \setminus \mathcal{I}$, while the ideal $\mathcal{I} = (\mathbf{x}_1^2 + \mathbf{x}_2^2)$ is radical but not real radical since $\mathbf{x}_1, \mathbf{x}_2 \in \sqrt[\mathbb{R}]{\mathcal{I}} \setminus \mathcal{I}$. The following celebrated results relate (real) radical and vanishing ideals.

Theorem 2.1 ([16, 52, 98]). *Let $\mathcal{I}$ be an ideal in $\mathbb{R}[\mathbf{x}]$. Then, $\sqrt{\mathcal{I}} = \mathcal{I}(V(\mathcal{I}))$* **(Hilbert's Nullstellensatz)** *and $\sqrt[\mathbb{R}]{\mathcal{I}} = \mathcal{I}(V_{\mathbb{R}}(\mathcal{I}))$* **(Real Nullstellensatz)**.

As $\mathcal{I} \subseteq \mathcal{I}(V(\mathcal{I})) \subseteq \mathcal{I}(V_{\mathbb{R}}(\mathcal{I}))$, $\mathcal{I}$ real radical implies $\mathcal{I}$ radical and, moreover, $V(\mathcal{I}) = V_{\mathbb{R}}(\mathcal{I}) \subseteq \mathbb{R}^n$ if the real variety $V_{\mathbb{R}}(\mathcal{I})$ is finite. Moreover, an ideal $\mathcal{I}$ is zero-dimensional precisely when $V(\mathcal{I})$ is finite. Then there is a well known relationship between the cardinality of the variety $V(\mathcal{I})$ and the dimension of the quotient space $\mathbb{R}[\mathbf{x}]/\mathcal{I}$ (see e.g. [16]).

Proposition 2.2. *An ideal $\mathcal{I}$ in $\mathbb{R}[\mathbf{x}]$ is zero-dimensional (i.e., the variety $V(\mathcal{I})$ is finite) if and only if the vector space $\mathbb{R}[\mathbf{x}]/\mathcal{I}$ is finite dimensional. Moreover, we have the inequality: $|V(I)| \leq \dim \mathbb{R}[\mathbf{x}]/\mathcal{I}$, with equality if and only if the ideal $\mathcal{I}$ is radical.*

The eigenvalue method for computing the variety $V(\mathcal{I})$. We now recall how to find the variety $V(\mathcal{I})$ of a zero-dimensional ideal $\mathcal{I}$ by computing the eigenvalues of the multiplication operator in the quotient algebra $\mathbb{R}[\mathbf{x}]/\mathcal{I}$, since this technique is used for finding the global minimizers of problem (P) (see [44]). Given a polynomial $h \in \mathbb{R}[\mathbf{x}]$, consider the 'multiplication by h' linear map in $\mathbb{R}[\mathbf{x}]/\mathcal{I}$:

$$\begin{aligned} m_h : \mathbb{R}[\mathbf{x}]/\mathcal{I} &\longrightarrow \mathbb{R}[\mathbf{x}]/\mathcal{I} \\ f + \mathcal{I} &\longmapsto fh + \mathcal{I} \end{aligned}$$

and let M_h denote its matrix in a given linear basis $\mathcal{B} = \{b_1, \ldots, b_N\}$ of $\mathbb{R}[\mathbf{x}]/\mathcal{I}$.

Theorem 2.3. *Assume $N = \dim \mathbb{R}[\mathbf{x}]/\mathcal{I} < \infty$, let $\mathcal{B} = \{b_1, \ldots, b_N\}$ be a linear basis of $\mathbb{R}[\mathbf{x}]/\mathcal{I}$, and let $[v]_{\mathcal{B}} = (b_1(v), \ldots, b_N(v))^{\mathsf{T}}$ be the vector of evaluations at $v \in V(\mathcal{I})$ of the polynomials in $\mathcal{B}$. For any $h \in \mathbb{R}[\mathbf{x}]$, the eigenvalues of the multiplication matrix M_h are the evaluations $h(v)$ of h at the points $v \in V(\mathcal{I})$, with corresponding (left) eigenvectors $[v]_{\mathcal{B}}$. That is, $M_h^{\mathsf{T}}[v]_{\mathcal{B}} = h(v)[v]_{\mathcal{B}}$ for all $v \in V(\mathcal{I})$.*

If $\mathcal{I}$ is radical then $|V(\mathcal{I})| = N$ (by Proposition 2.2) and the matrix M_h has a full set of linearly independent eigenvectors ($[v]_{\mathcal{B}}$ for $v \in V(\mathcal{I})$). These vectors can be found by

computing the eigenvalues of M_h^{T} (assuming the values $h(v)$ are pairwise distinct which can be achieved e.g. by selecting a random linear polynomial h) and it is then easy to recover the points $v \in V(\mathcal{I})$ from these vectors $[v]_{\mathcal{B}}$.

We illustrate this method applied to the univariate case. Say $\mathcal{I} = (p)$, where p is the polynomial: $p = \mathbf{x}^d - p_{d-1}\mathbf{x}^{d-1} - \ldots - p_0$. The set $\mathcal{B} = \{1, \mathbf{x}, \ldots, \mathbf{x}^{d-1}\}$ is a basis of $\mathbb{R}[\mathbf{x}]/(p)$ and with respect to this basis the 'multiplication by $\mathbf{x}$' matrix has the form

$$M_{\mathbf{x}} = \begin{pmatrix} 0 & \ldots & 0 & p_0 \\ & I_{d-1} & & \vdots \\ & & & p_{d-1} \end{pmatrix}.$$

Its characteristic polynomial is $\det(M_{\mathbf{x}} - tI) = (-1)^d p(t)$, hence the eigenvalues of the matrix $M_{\mathbf{x}}$ are the roots of p and indeed $M_{\mathbf{x}}^{\mathsf{T}}[v]_{\mathcal{B}} = v[v]_{\mathcal{B}}$ holds if $p(v) = 0$.

Semidefinite optimization. $\mathcal{S}^n$ is the set of real symmetric $n \times n$ matrices, equipped with the trace inner product $\langle X, Y\rangle = \mathrm{Tr}(X^{\mathsf{T}}Y) = \sum_{i,j=1}^n X_{ij}Y_{ij}$. The notation $X \succeq 0$ means that X is positive semidefinite (i.e., $x^{\mathsf{T}}Xx \geq 0$ for all $x \in \mathbb{R}^n$) and $\mathcal{S}_+^n \subseteq \mathcal{S}^n$ is the cone of positive semidefinite matrices. The cone $\mathcal{S}_+^n$ is self-dual: $X \in \mathcal{S}^n$ is positive semidefinite if and only if $\langle X, Y\rangle \geq 0$ for all $Y \in \mathcal{S}_+^n$.

Given matrices $C, A_1, \ldots, A_m \in \mathcal{S}^n$ and a vector $b \in \mathbb{R}^m$, a *semidefinite program* in standard primal form and its *dual semidefinite program* read:

$$p^* = \sup_{X \in \mathcal{S}^n} \{\langle C, X\rangle : \langle A_j, X\rangle = b_j \ (j \in [m]),\ X \succeq 0\}, \tag{P-SDP}$$

$$d^* = \inf_{y \in \mathbb{R}^m} \{b^{\mathsf{T}}y : \sum_{j=1}^m y_j A_j - C \succeq 0\}. \tag{D-SDP}$$

Weak duality holds: $p^* \leq d^*$ (since $X, Y = \sum_{j=1}^m y_j A_j - C \succeq 0$ implies $\langle X, Y\rangle \geq 0$). Moreover, if (P-SDP) is bounded and has a positive definite feasible solution X, then strong duality holds: $p^* = d^*$. Semidefinite programs can be solved (approximatively) in polynomial time, using the ellipsoid method (since one can test if a rational matrix is positive semidefinite using Gaussian elimination). However, the ellipsoid method is not efficient in practice, and efficient algorithms used in practical implementations rely on interior-point algorithms. (See e.g. [5, 21, 99, 100].) On the other hand, the exact complexity is not known of the problem of *testing feasibility* of a semidefinite program: given integral matrices $C, A_1, \ldots, A_m \in \mathcal{S}^n$,

$$\text{decide whether there exists } y \in \mathbb{R}^n \text{ such that } C + y_1A_1 + \ldots + y_mA_m \succeq 0. \tag{F}$$

An obvious difficulty is that there might be only irrational solutions. It is known that (F) belongs to NP if and only if it belongs to co-NP ([88], see also [51]). Moreover, (F) can be solved in polynomial time when fixing either m or n [46] and, when fixing m, one can check in polynomial time if (F) has a rational solution [46].

Recognizing sums of squares of polynomials. It turns out that checking whether a polynomial $f = \sum_{\alpha \in \mathbb{N}_{2t}^n} f_\alpha \mathbf{x}^\alpha$ can be written as a sum of squares of polynomials amounts to checking whether the following semidefinite program:

$$\sum_{\beta,\gamma \in \mathbb{N}_t^n : \beta+\gamma=\alpha} X_{\beta,\gamma} = f_\alpha \quad (\alpha \in \mathbb{N}_{2t}^n),\ \ X \succeq 0, \tag{2.1}$$

(in the matrix variable $X = (X_{\beta,\gamma})_{\beta,\gamma\in\mathbb{N}^n_t}$) admits a feasible solution. To see this, assume $f = \sum_{j=1}^k p_j^2$. Then each p_j has degree at most t and can be written as $p_j = \sum_\alpha (p_j)_\alpha \mathbf{x}^\alpha = \mathbf{p_j}^\mathsf{T}[\mathbf{x}]_t$, where $\mathbf{p_j} = ((p_j)_\alpha)$ is the vector of coefficients of p_j in the monomial basis. Therefore, $f = \sum_{j=1}^k p_j^2 = [\mathbf{x}]_t^\mathsf{T}(\sum_{j=1}^k \mathbf{p_j}\mathbf{p_j}^\mathsf{T})[\mathbf{x}]_t = [\mathbf{x}]_t^\mathsf{T} P[\mathbf{x}]_t$, where the matrix $P = \sum_{j=1}^k \mathbf{p_j}\mathbf{p_j}^\mathsf{T}$ is positive semidefinite. Moreover, by equating the coefficients of both polynomials f and $[\mathbf{x}]_d^\mathsf{T} P[\mathbf{x}]_d$ in the identity $f = [\mathbf{x}]_t^\mathsf{T} P[\mathbf{x}]_t$, it follows that P satisfies the system (2.1). The argument can be easily reversed: any feasible solution of (2.1) gives rise to a sum of squares decomposition of f.

More generally, given polynomials $f, g_1, \ldots, g_m \in \mathbb{R}[\mathbf{x}]$, the problem of finding a decomposition of the form $f = \sigma_0 + \sigma_1 g_1 + \ldots + \sigma_m g_m$, where $\sigma_0, \sigma_1, \ldots, \sigma_m$ are sums of squares with a given degree bound: $\deg(\sigma_0), \deg(\sigma_j g_j) \le 2t$, can also be cast as a semidefinite program. This program is analogue to (2.1), but it now involves $m+1$ positive semidefinite matrices $X_0, X_1, \ldots, X_m$, where X_0 is indexed by $\mathbb{N}^n_t$ (corresponding to σ_0) and X_j by $\mathbb{N}^n_{t-\lceil \deg(g_j)/2\rceil}$ (corresponding to σ_j). Of course one should adequately define the affine constraints in the semidefinite program.

3. Positive polynomials and sums of squares

3.1. Positivity certificates. Understanding the link between positive polynomials and sums of squares is a classic question which goes back to work of Hilbert around 1890. Hilbert realized that not every nonnegative polynomial is a sum of squares of polynomials and he characterized when this happens.

Theorem 3.1 (Hilbert [45]). *Every nonnegative polynomial of degree $2d$ in n variables is a sum of squares of polynomials if and only if we are in one of the following three cases:* $(n = 1, 2d)$, $(n, 2d = 2)$, *and* $(n = 2, 2d = 4)$.

In all other cases, Hilbert showed the existence of a nonnegative polynomial which is not sos. The first explicit construction was found only sixty years later by Motzkin: the *Motzkin polynomial* $M = \mathbf{x}_1^2\mathbf{x}_2^2(\mathbf{x}_1^2 + \mathbf{x}_2^2 - 3) + 1$ is nonnegative but not a sum of squares of polynomials. However, the polynomial $(1 + \mathbf{x}_1^2 + \mathbf{x}_2^2)M$ is a sum of squares of polynomials, which certifies the positivity of M. We refer to [89] for an historic account and for more examples. We also refer to [7] for an in-depth study of the two smallest cases $(n = 2, 2d = 6)$ and $(n = 3, 2d = 4)$ when not all nonnegative polynomials are sums of squares.

If we are not in one of the special three cases of Theorem 3.1, then the inclusion $\Sigma[\mathbf{x}]_{2d} \subseteq \mathcal{P}(\mathbb{R}^n) \cap \mathbb{R}[\mathbf{x}]_{2d}$ is strict. Are these two sets far apart or not? That is, are there few or many sums of squares within nonnegative polynomials? The answer depends whether the degree and the number of variables are fixed or not.

On the one hand, sums of squares are dense within nonnegative polynomials if we allow the degree to grow. Lasserre and Netzer [60] show the following explicit sums of squares approximation: if f is nonnegative over the box $[-1, 1]^n$ then for any $\epsilon > 0$ there exists $k \in \mathbb{N}$ such that the perturbed polynomial $f + \epsilon(1 + \sum_{i=1}^n \mathbf{x}_i^{2k})$ is a sum of squares of polynomials. (See also Lasserre [55]).

On the other hand, if we fix the degree but let the number of variables grow, then there are significantly more nonnegative polynomials than sums of squares: Blekherman [6] shows

that the ratio of volumes of (sections of) the cone of sums of squares and the cone of nonnegative polynomials tends to 0 as n goes to ∞.

At the 1900 International Congress of Mathematicians in Paris, Hilbert asked whether every nonnegative polynomial can be written as a sum of squares of *rational* functions. This question, known as Hilbert's 17th problem, was answered in the affirmative in 1927 by Artin, whose work led the foundations of the field of real algebraic geometry.

Sums of squares certificates (also known as *Positivstellensätze*) are known for characterizing positivity over a general basic closed semialgebraic set K of the form (1.1). They involve weighted combinations of the polynomials $g_1, \ldots, g_m$ describing the set K. The *quadratic module* generated by $g = (g_1, \ldots, g_m)$ is the set

$$\mathcal{Q}(g) = \{\sigma_0 + \sigma_1 g_1 + \ldots + \sigma_m g_m : \sigma_0, \ldots, \sigma_m \in \Sigma[\mathbf{x}]\}, \tag{3.1}$$

the *truncated quadratic module* $\mathcal{Q}_t(g)$ is its subset obtained by restricting the degrees: $\deg(\sigma_j g_j) \leq 2t$ (setting $g_0 = 1$), and the *preordering* $\mathcal{T}(g)$ is the quadratic module generated by the 2^m polynomials $g^e = g_1^{e_1} \cdots g_m^{e_m}$ for $e \in \{0,1\}^m$.

Theorem 3.2 (Krivine [52], Stengle [98]). *Let $f \in \mathbb{R}[\mathbf{x}]$ and K be as in (1.1).*

(i) *$f > 0$ on K if and only if $fq = 1 + p$ for some $p, q \in \mathcal{T}(g)$.*

(ii) *$f \geq 0$ on K if and only if $fq = f^{2k} + p$ for some $p, q \in \mathcal{T}(g)$ and $k \in \mathbb{N}$.*

(iii) *$f = 0$ on K if and only if $-f^{2k} \in \mathcal{T}(g)$ for some $k \in \mathbb{N}$.*

In each case it is clear that the 'if part' gives a certificate that f is positive (nonnegative, or vanishes) on K, the hard part is showing the existence of such a certificate. These certificates use polynomials in $\mathcal{T}(g)$ and thus they can be checked with semidefinite optimization, once a bound on the degrees has been set. However they are not directly useful for our polynomial optimization problem (P). Indeed, in view of Theorem 3.2 (i), one would need to search for the largest scalar λ for which there exist $p, q \in \mathcal{T}(g)$ such that $(f - \lambda)q = 1 + p$, thus involving a quadratic term λq which cannot be dealt with directly using semidefinite optimization.

To go around this difficulty one may instead use the simpler "denominator free" positivity certificates of Schmüdgen and Putinar, which hold in the case when the semialgebraic set K is compact. The following condition:

$$\exists R > 0 \text{ such that } R - \mathbf{x}_1^2 - \ldots - \mathbf{x}_n^2 \in \mathcal{Q}(g), \tag{A}$$

known as the *Archimedean condition*, allows easier positivity certificates using the quadratic module $\mathcal{Q}(g)$. Note that K is compact if (A) holds.

Theorem 3.3 (Schmüdgen [92]). *Assume that the set K in (1.1) is compact. If the polynomial f is positive on K (i.e., $f(x) > 0$ for all $x \in K$), then $f \in \mathcal{T}(g)$.*

Theorem 3.4 (Putinar [86]). *Assume that the Archimedean condition (A) holds. If the polynomial f is positive on K, then $f \in \mathcal{Q}(g)$.*

3.2. Semidefinite relaxations for (P). Motivated by Putinar's result, Lasserre [53] introduced the following relaxations for the polynomial optimization problem (P). For any integer $t \geq d_f = \lceil \deg(f)/2 \rceil$, consider the parameters

$$f_t^{\text{sos}} = \sup_{\lambda \in \mathbb{R}} \{\lambda : f - \lambda \in \mathcal{Q}_t(g)\}, \tag{SOSt}$$

which form a monotone nondecreasing sequence: $f_t^{\rm sos} \le f_{t+1}^{\rm sos} \le \ldots \le f_{\rm min}$.

Each program (SOSt) can be written as a semidefinite program (recall Section 2). Moreover, the dual semidefinite program can be expressed as follows:

$$f_t^{\rm mom} = \inf_{L \in \mathbb{R}[\mathbf{x}]_{2t}^*} \{L(f) : L(f) = 1,\ L(p) \ge 0\ \forall p \in \mathcal{Q}_t(g)\}, \tag{MOMt}$$

where $\mathbb{R}[\mathbf{x}]_{2t}^*$ denotes the set of linear functionals on $\mathbb{R}[\mathbf{x}]_{2t}$. The parameters $f_{\rm min}$, $f_t^{\rm sos}$ and $f_t^{\rm mom}$ satisfy:

$$f_t^{\rm sos} \le f_t^{\rm mom} \le f_{\rm min}. \tag{3.2}$$

The inequality $f_t^{\rm sos} \le f_t^{\rm mom}$ is easy (by weak duality) and $f_t^{\rm mom} \le f_{\rm min}$ is explained below in Section 4.1. There is no duality gap: $f_t^{\rm sos} = f_t^{\rm mom}$, for instance if the set K has an interior point. In the compact case the asymptotic convergence of the bounds to the infimum of f is guaranteed by Putinar's theorem.

Theorem 3.5. *(Lasserre [53]) Assume that assumption (A) holds (and thus K is compact). Then,* $\lim_{t\to\infty} f_t^{sos} = \lim_{t\to\infty} f_t^{mom} = f_{min}$.

Proof. For any $\epsilon > 0$, the polynomial $f - f_{\rm min} + \epsilon$ is positive on K and thus, by Theorem 3.4, it belongs to $\mathcal{Q}_t(g)$ for some t, which implies $f_t^{\rm sos} \ge f_{\rm min} - \epsilon$. □

In order to discuss further properties of the dual (moment) programs (MOMt), we need to go in some detail about the moment problem. This is what we do in the next sections and we come back to the hierarchies later in Section 4.4.

4. Moment sequences and moment matrices

4.1. The moment problem. Given a (positive Borel) measure μ on a set $K \subseteq \mathbb{R}^n$, consider the linear functional $L_\mu \in \mathbb{R}[\mathbf{x}]^*$ defined by

$$L_\mu(f) = \int_K f(x) d\mu = \sum_\alpha f_\alpha \left(\int_K x^\alpha d\mu \right) \quad \text{for } f \in \mathbb{R}[\mathbf{x}], \tag{4.1}$$

which thus depends linearly on the moments $\int_K x^\alpha d\mu$ of the measure μ. The classical moment problem asks to characterize the linear functionals $L \in \mathbb{R}[\mathbf{x}]^*$ admitting such a representing measure μ, i.e., being of the form $L = L_\mu$. The following result (due to Haviland) makes the link to polynomial positivity: $L = L_\mu$ for some measure μ on K if and only if L is nonnegative on $\mathcal{P}(K)$.

Let us go back to problem (P). Following Lasserre [53], we observe that the infimum of f over the set K can be reformulated as

$$f_{\rm min} = \inf_\mu \{L_\mu(f) : \mu \text{ is a probability measure on } K\}.$$

Indeed, as $f(x) \ge f_{\rm min}$ for all $x \in K$, by integrating both sides over K for an arbitrary probability measure μ on K, we obtain that $L_\mu(f) \ge f_{\rm min}$. For the reverse inequality, choose μ to be the Dirac measure at an arbitrary point $x \in K$, so that $L_\mu(f) = f(x)$ and thus $\inf_\mu L_\mu(f) \le f(x)$.

If μ is a probability measure on K, then L_μ is nonnegative on $\mathcal{P}(K)$ and thus on its subset $\mathcal{Q}_t(g)$, which implies the inequality $f_t^{\text{mom}} \le f_{\min}$ from (3.2). Moreover, the relaxation (MOMt) is exact, i.e., $f_t^{\text{mom}} = f_{\min}$, if it has an optimal solution of the form L_μ where μ is a probability measure on K. This observation motivates searching for sufficient conditions for existence of a representing measure. This is treated in the rest of the section.

If $L \in \mathbb{R}[\mathbf{x}]^*$ has a representing measure then L must be nonnegative on $\mathcal{P}(K)$ and thus on the subcone $\Sigma[\mathbf{x}]$ of all sums of squares. The nonnegativity condition of L over $\Sigma[\mathbf{x}]$ can be conveniently expressed using the following 'Hankel type' matrix $M(L)$:

$$M(L) = (L(\mathbf{x}^\alpha \mathbf{x}^\beta))_{\alpha,\beta\in\mathbb{N}^n},$$

which is indexed by $\mathbb{N}^n$ and called the *moment matrix* of L.

Indeed, note that $L(pq) = \mathbf{p}^{\mathsf{T}} M(L)\mathbf{q}$ for any $p, q \in \mathbb{R}[\mathbf{x}]$. Therefore, L is nonnegative over $\Sigma[\mathbf{x}]$ if and only if $M(L) \succeq 0$. Moreover, for $g \in \mathbb{R}[\mathbf{x}]$, L is nonnegative on the set $g\Sigma[\mathbf{x}] = \{g\sigma : \sigma \in \Sigma[\mathbf{x}]\}$ if and only if $M(gL) \succeq 0$, where $gL \in \mathbb{R}[\mathbf{x}]^*$ is the new linear functional defined by $(gL)(p) = L(gp)$ for $p \in \mathbb{R}[\mathbf{x}]$.

For example, in the univariate case, L has a representing measure on $\mathbb{R}$ if and only if $M(L) \succeq 0$ (Hamburger's theorem), L has a representing measure on $\mathbb{R}_+$ if and only if $M(L), M(\mathbf{x}L) \succeq 0$ (Stieltjes' theorem), and L has a representing measure on $[0, 1]$ if and only if $M(\mathbf{x}L), M((1-\mathbf{x})L) \succeq 0$ (Hausdorff's theorem).

Both Theorems 3.3-3.4 have counterparts for the moment problem. If K is compact, then L has a representing measure on K if and only if $L \ge 0$ on $\mathcal{T}(g)$ (Schmüdgen [92]) or, equivalently, $L \ge 0$ on $\mathcal{Q}(g)$ if (A) holds (Putinar [86]).

4.2. Finite rank moment matrices. As we saw above, a necessary condition for $L \in \mathbb{R}[\mathbf{x}]^*$ to have a representing measure is positive semidefiniteness of its moment matrix. Although not sufficient in general, it turns out that this condition is sufficient in the case when $M(L)$ has finite rank ([17], see Theorem 4.1 below). As this result plays a crucial role for studying the finite convergence of the relaxations (MOMt) for (P), we discuss it in detail.

In what follows, $\operatorname{Ker} M(L)$ denotes the kernel of $M(L)$, which consists of the polynomials $p \in \mathbb{R}[\mathbf{x}]$ for which $L(pq) = 0$ for all $q \in \mathbb{R}[\mathbf{x}]$. Hence $\operatorname{Ker} M(L)$ is an ideal in $\mathbb{R}[\mathbf{x}]$. Moreover, $\operatorname{Ker} M(L)$ is real radical if $M(L) \succeq 0$ (since, when $M(L) \succeq 0$, a polynomial p belongs to $\operatorname{Ker} M(L)$ if and only if $L(p^2) = 0$).

Consider a measure μ and the corresponding linear functional L_μ as in (4.1). Its support is contained in the real variety of the polynomials in the kernel of $M(L_\mu)$: $\operatorname{Supp}(\mu) \subseteq V_{\mathbb{R}}(\operatorname{Ker} M(L_\mu))$. When $\mu = \delta_v$ is the Dirac measure at a point $v \in \mathbb{R}^n$, L_μ is the *evaluation* L_v at v, defined by $L_v(p) = p(v)$ for all $p \in \mathbb{R}[\mathbf{x}]$. If the suppport of μ is finite (i.e., μ is *finite atomic*), say $\operatorname{Supp}(\mu) = \{v_1, \ldots, v_r\}$, then L_μ is a conic combination of evaluations at the v_i's: $L_\mu = \sum_{i=1}^r \lambda_i L_{v_i}$ for some scalars $\lambda_i > 0$. The following theorem shows that this describes all the linear functionals $L \in \mathbb{R}[\mathbf{x}]^*$ with $M(L) \succeq 0$ and $\operatorname{rank} M(L) < \infty$. We present our simple real algebraic proof from [64] (see also [68]).

Theorem 4.1. *(Curto and Fialkow [17]) Let $L \in \mathbb{R}[\mathbf{x}]^*$. Assume that $M(L) \succeq 0$ and that $M(L)$ has finite rank r. Then L has a (unique) representing measure μ. Moreover, μ is finite atomic with r atoms and supported by $V(\operatorname{Ker} M(L))$.*

Proof. As $M(L) \succeq 0$, its kernel $\mathcal{I} := \operatorname{Ker} M(L)$ is a real radical ideal in $\mathbb{R}[\mathbf{x}]$.

Moreover, the quotient space $\mathbb{R}[\mathbf{x}]/\mathcal{I}$ has finite dimension r. This is because we have: $\operatorname{rank} M(L) = r$ and any set of monomials $\mathcal{B}$ indexing a maximal linearly independent set of columns of $M(L)$ is also maximal linearly independent in $\mathbb{R}[\mathbf{x}]/\mathcal{I}$.

Applying Proposition 2.2, we can conclude that the variety of the ideal $\mathcal{I}$ is contained in $\mathbb{R}^n$ and has cardinality r. Set $V(\mathcal{I}) = \{v_1, \ldots, v_r\} \subseteq \mathbb{R}^n$.

We consider interpolation polynomials $p_{v_1}, \ldots, p_{v_r} \in \mathbb{R}[\mathbf{x}]$ at the points of $V(\mathcal{I})$, i.e., satisfying $p_{v_i}(v_j) = \delta_{i,j}$. As the polynomial $p_{v_i} - p_{v_i}^2$ vanishes on the variety $V(\mathcal{I})$, it belongs to the ideal $\mathcal{I}(V(\mathcal{I}))$, which is equal to $\mathcal{I}$ (since $\mathcal{I}$ is real radical). Hence, $L(p_{v_i}) = L(p_{v_i}^2)$, since $p_{v_i} - p_{v_i}^2 \in \mathcal{I} = \operatorname{Ker} M(L)$. Moreover, $L(p_{v_i}^2) \geq 0$ since $M(L) \succeq 0$. Furthermore, $L(p_{v_i}^2) \neq 0$, since otherwise p_{v_i} would belong to $\operatorname{Ker} M(L)$ and thus it would vanish at v_i, a contradiction.

We now claim that $L = \sum_{i=1}^r L(p_{v_i}) L_{v_i}$. Indeed, any $p \in \mathbb{R}[\mathbf{x}]$ can be written as $p = \sum_{i=1}^r p(v_i) p_{v_i} + q$, where $q \in \mathcal{I}$. Hence, $L(q) = 0$ and thus $L(p) = \sum_{i=1}^r p(v_i) L(p_{v_i}) = \sum_{i=1}^r L_{v_i}(p) L(p_{v_i})$. Hence we have shown that L has a finite r-atomic representing measure: $\mu = \sum_{i=1}^r L(p_{v_i}) \delta_{v_i}$, which concludes the proof. □

4.3. Flat extensions of truncated moment matrices. To make the link with the relaxations (MOMt) for problem (P), we introduce the *truncated moment matrix* of $L \in \mathbb{R}[\mathbf{x}]_{2t}^*$, which is the following matrix indexed by $\mathbb{N}_t^n$:

$$M_t(L) = (L(\mathbf{x}^\alpha \mathbf{x}^\beta))_{\alpha,\beta \in \mathbb{N}_t^n}.$$

Following Curto and Fialkow [17] we say that $M_t(L)$ is a *flat extension* of (its principal submatrix) $M_{t-1}(L)$ if

$$\operatorname{rank} M_t(L) = \operatorname{rank} M_{t-1}(L). \tag{4.2}$$

The following result claims that any such moment matrix can be extended to an infinite moment matrix of the same rank.

Theorem 4.2 ([17]). *Let $L \in \mathbb{R}[\mathbf{x}]_{2t}^*$. If $M_t(L)$ is a flat extension of $M_{t-1}(L)$, i.e., (4.2) holds, then there exists $\tilde{L} \in \mathbb{R}[\mathbf{x}]^*$ which extends L (i.e., $L = \tilde{L}$ on $\mathbb{R}[\mathbf{x}]_{2t}$) and has the property that $M(\tilde{L})$ is a flat extension of $M_t(L)$:* $\operatorname{rank} M(\tilde{L}) = \operatorname{rank} M_t(L)$.

The proof is elementary, exploiting the fact that the kernel of $M(\tilde{L})$ is an ideal. Indeed the relations expressing the monomials of degree t in terms of polynomials of degree at most $t-1$ (modulo the kernel of $M_t(L)$) can be used to express recursively any monomial of degree at least $t+1$ in terms of polynomials of degree at most t (modulo the ideal generated by the kernel of $M_t(L)$). Combining Theorems 4.1 and 4.2, we arrive at the following result.

Theorem 4.3. *Let $L \in \mathbb{R}[\mathbf{x}]_{2t}^*$ and assume that $M_t(L) \succeq 0$ and (4.2) holds. Then L has a finite atomic representing measure μ, whose support is given by the variety of the kernel of $M_t(L)$: $V(\operatorname{Ker} M_t(L)) = \operatorname{Supp}(\mu) \subseteq \mathbb{R}^n$. Moreover, the ideal generated by the kernel of $M_t(L)$ is equal to the kernel of $M(L_\mu)$: $(\operatorname{Ker} M_t(L)) = \operatorname{Ker} M(L_\mu)$, and it is a real radical ideal.*

To be able to claim that the representing measure μ is supported within a given semialgebraic set K like (1.1), it suffices to add the *localizing conditions* $M_{t-d_{g_j}}(g_j L) \succeq 0$ (for $j \in [m]$), where g_j are the polynomials defining K and $d_{g_j} = \lceil \deg(g_j)/2 \rceil$, and to assume a stronger flatness condition:

$$\operatorname{rank} M_t(L) = \operatorname{rank} M_{t-d_K}(L), \quad \text{where } d_K = \max\{d_{g_j} : j \in [m]\}. \tag{4.3}$$

Theorem 4.4 ([18]). *Assume $L \in \mathbb{R}[\mathbf{x}]_{2t}^*$ satisfies $M_t(L) \succeq 0$, $M_{t-d_{g_j}}(g_j L) \succeq 0$ for $j \in [m]$, and the flatness condition (4.3). Then L has a representing measure whose support is contained in the set K.*

Proof. We give our simple proof from [64]. We already know that L has a representing measure μ with $\mathrm{Supp}(\mu) =: \{v_1, \ldots, v_r\} \subseteq \mathbb{R}^n$, where $r = \mathrm{rank} M_t(L)$ and $L = \sum_{i=1}^r \lambda_i L_{v_i}$ with $\lambda_i = L(p_{v_i}) > 0$. It suffices now to show that each point $v_i \in \mathrm{Supp}(\mu)$ belongs to K, i.e., that $g_j(v_i) \geq 0$ for all $j \in [m]$. For this, the simple but crucial observation is that we can choose the interpolation polynomials p_{v_i} at the v_i's in such a way that they all have degree at most $t - d_K$ (which follows using condition (4.3)). As each polynomial p_{v_i} has degree at most $t - d_K \leq t - d_{g_j}$ and $M_{t-d_{g_j}}(g_j L) \succeq 0$, we can conclude that $0 \leq (g_j L)(p_{v_i}^2) = L(p_{v_i}^2 g_j)$, which implies directly that $g_j(v_i) \geq 0$. □

4.4. The moment relaxations for (P). We now return to the moment relaxation (MOMt) for problem (P) introduced earlier in Section 3.2. First, using truncated moment matrices, it can be reformulated as follows:

$$f_t^{\mathrm{mom}} = \inf_{L \in \mathbb{R}[\mathbf{x}]_{2t}^*} \{L(f) : L(1) = 1,\ M_t(L) \succeq 0,\ M_{t-d_{g_j}}(g_j L) \succeq 0\ (j \in [m])\}, \quad \text{(MOMt)}$$

(explaining the name 'moment' and the notation 'f_t^{mom}'). Recall that $f_t^{\mathrm{mom}} \leq f_{\min}$ from (3.2). Using the preceding results about flat extensions of moment matrices, we can now present the following **optimality certificate** for the relaxation (MOMt), which permits to claim that the infimum of f is reached: $f_t^{\mathrm{mom}} = f_{\min}$.

Theorem 4.5. *Let K_f denote the set of global minimizers of problem (P) and set $d_f = \lceil \deg(f)/2 \rceil$, $d_{g_j} = \lceil \deg(g_j)/2 \rceil$, $d_K = \max\{d_{g_j} : j \in [m]\}$. Let $L \in \mathbb{R}[\mathbf{x}]_{2t}^*$ be an optimal solution of the program (MOMt). Assume that L satisfies the following flatness condition:*

$$\mathrm{rank} M_s(L) = \mathrm{rank} M_{s-d_K}(L) \text{ for some } s \text{ satisfying } \max\{d_f, d_K\} \leq s \leq t. \quad (4.4)$$

Then, $f_t^{mom} = f_{min}$ and $V(\mathrm{Ker}\, M_s(L)) \subseteq K_f$. Moreover, if $\mathrm{rank} M_s(L)$ is maximum among all optimal solutions of (MOMt), then equality: $V(\mathrm{Ker}\, M_s(L)) = K_f$ holds and $\mathcal{I}(K_f) = (\mathrm{Ker}\, M_s(L))$.

Proof. Assume $s = t$ (to simplify notation). By Theorem 4.4, L has a representing measure μ with $\mathrm{Supp}(\mu) \subseteq K$. That is, $L = \sum_{i=1}^r \lambda_i L_{v_i}$, where $\lambda_i > 0$, $\sum_i \lambda_i = 1$, and $\{v_1, \ldots, v_r\} = V(\mathrm{Ker}\, M_t(L)) \subseteq K$. Then, $f_t^{\mathrm{mom}} = L(f) = \sum_{i=1}^r \lambda_i f(v_i) \geq f_{\min}$. This implies equality $f_t^{\mathrm{mom}} = f_{\min}$ and $f(v_i) = f_{\min}$ for all $i \in [r]$, and thus we can conclude that $V(\mathrm{Ker}\, M_t(L)) = \{v_1, \ldots, v_r\} \subseteq K_f$.

Assume now that $M_t(L)$ has maximum rank among the optimal solutions of (MOMt). As the evaluation L_v at any point $v \in K_f$ is also an optimal solution of (MOMt), we deduce that $\mathrm{rank}\, M_t(L_v) \leq \mathrm{rank}\, M_t(L)$, which implies that $\mathrm{Ker}\, M_t(L) \subseteq \mathrm{Ker}\, M_t(L_v) \subseteq \mathcal{I}(v)$ for all $v \in K_f$. Hence, $\mathrm{Ker}\, M_t(L)$ is contained in $\cap_{v \in K_f} \mathcal{I}(v) = \mathcal{I}(K_f)$. By taking the varieties on both sides, we obtain that $K_f \subseteq V(\mathrm{Ker}\, M_t(L))$, which implies $K_f = V(\mathrm{Ker}\, M_t(L))$ and thus $\mathcal{I}(K_f) = (\mathrm{Ker}\, M_s(L))$ (since $(\mathrm{Ker}\, M_t(L))$ is real radical by Theorem 4.3). □

The above result is the theoretical core of the moment approach for problem (P). It has been implemented in the numerical algorithm GloptiPoly. There are several other implementations of the sos/moment approach, including SOSTOOLS, YALMIP, and SparsePOP (tuned to exploit sparsity structure). We conclude with some comments and pointers to a few additional results from the growing literature.

- *The maximality assumption on the rank of the optimal solution is not restrictive.* On the contrary, most interior point algorithms currently used to solve semidefinite programs return an optimal solution lying in the relative interior of the optimal face and thus one with maximum possible rank (see [21]).
- *Under the assumptions of Theorem 4.5, problem (P) has finitely many global minimizers and they can be found using the eigenvalue method from Section 2.* Indeed, we know that the set of global minimizers is $K_f = V(\operatorname{Ker} M_s(L))$ and that the quotient space $\mathbb{R}[\mathbf{x}]/(\operatorname{Ker} M_s(L))$ has dimension $\operatorname{rank} M_s(L) = \operatorname{rank} M_{s-d_K}(L)$. Hence any set of monomials indexing a maximal linearly independent set of columns of the matrix $M_{t-d_K}(L)$ is a linear basis of $\mathbb{R}[\mathbf{x}]/(\operatorname{Ker} M_s(L))$. So we can construct the multiplication matrices in $\mathbb{R}[\mathbf{x}]/(\operatorname{Ker} M_s(L))$ and their eigenvalues/eigenvectors permit to extract the points in $V(\operatorname{Ker} M_s(L)) = K_f$.
- The flatness condition (4.4) can be used as a *concrete optimality stopping criterion*: if it is satisfied at a certain order t then the relaxation is exact and the algorithm stops after returning the infimum $f_{\min}$ and the set K_f of global minimizers. Otherwise one may compute the next relaxation of order $t+1$.
- In general, information about the global minimizers can be gained asymptotically from optimal solutions L^t to the relaxations (MOMt). In particular, if (P) has a unique minimizer x^*, then x^* can be found asymptotically as limit point as $t \to \infty$ of the sequences $(L^t(\mathbf{x}_1), \ldots, L^t(\mathbf{x}_n))$ [95]. See [77] for an extension to the case of finitely many global minimizers.

  In the compact case, the bounds f_t^{sos}, f_t^{mom} converge asymptotically to $f_{\min}$ (Theorem 3.5). What about **finite convergence**?
- By Theorem 4.5, the flatness condition (4.4) implies the finite convergence of the moment hierarchy (MOMt). Conversely, if the set of global minimizers is nonempty and finite, *the flatness condition (4.4) is also necessary for finite convergence of (MOMt) under some genericity assumptions on the polynomials* f, g_j [77].
- Finite convergence holds in the case when the description of the set K involves some polynomial equations $g_1(x) = 0, \ldots, g_k(x) = 0$ which have finitely many common real roots (since the flatness condition holds) [66, 68, 78].
- Finite convergence also holds in the *convex case*, when $f, -g_1, \ldots, -g_m$ are convex, the set K has a Slater point x_0 (i.e., $g_j(x_0) > 0$ if g_j is not linear), and the Hessian of f is positive definite at the (unique) global minimizer [23].
- Nie [80] shows that, under the Archimedean condition (A), *the Lasserre hierarchy applied to problem (P) has finite convergence generically*. More precisely, finite convergence holds when the classic nonlinear optimality conditions (constraint qualification, strict complementarity, and second order sufficient condition) hold at all global minimizers, and these conditions are satisfied generically.
- Finally we refer to [81] for degree bounds and estimates on the quality of the moment/sos bounds (see [22] for refined results when K is the hypercube).

5. Application to real roots and real radical ideals

The above strategy for computing the global minimizers of (P) was developed and applied by Lasserre, Laurent and Rostalski [57] to the problem of computing the common real roots of a system of polynomial equations: $g_1(x) = 0, \ldots, g_k(x) = 0$.

Computing all *complex* roots is a well studied problem. Several methods exist, including symbolic-numeric methods, which combine symbolic tools (like Gröbner or border bases) with numerical linear algebra (like computing eigenvalues, or univariate root finding), and homotopy continuation methods. As there might be much less real roots than complex ones it is desirable to have methods able to extract directly the real roots without dealing with the complex nonreal ones. This is precisely the feature of the real algebraic method of [57], which can be summarized as follows.

Consider the following instance of (P):

$$\min\{0 : g_1(x) = 0, \ldots, g_k(x) = 0\},$$

which asks to minimize the zero polynomial on the real algebraic variety of the ideal $\mathcal{I} = (g_1, \ldots, g_k)$, so that the set of global minimizers is precisely $V_{\mathbb{R}}(\mathcal{I})$.

Consider the moment relaxations (MOMt) for this problem. [57] shows that the flatness condition (4.4) holds for t large enough, assuming that the set $V_{\mathbb{R}}(\mathcal{I})$ is finite. Hence, by Theorem 4.5, it follows that the real radical ideal of $\mathcal{I}$ is found: $\sqrt[\mathbb{R}]{\mathcal{I}} = (\mathrm{Ker}\, M_s(L))$ and that the variety $V_{\mathbb{R}}(\mathcal{I})$ can be computed using the eigenvalue method applied to the quotient space $\mathbb{R}[\mathbf{x}]/(\mathrm{Ker}\, M_s(L))$ (as explained in the previous section). The fact that the kernel of $M_s(L)$ generates the vanishing ideal of $V_{\mathbb{R}}(\mathcal{I})$ is crucial, since this is the key property which permits to filter out all complex nonreal roots.

We point out that the equality $\sqrt[\mathbb{R}]{\mathcal{I}} = (\mathrm{Ker}\, M_t(L))$ holds for t large enough, even if the variety $V_{\mathbb{R}}(\mathcal{I})$ is infinite. The difficulty, however, is to detect when one has reached such order t, since it is not clear how to detect it algorithmically (as the flatness condition cannot hold when the real variety is not finite).

We refer to [57, 58], [1, Chap.2] for details and extensions. The recent work [59] develops a *sparse* version of the moment method able to work with smaller matrices, indexed by smaller sets of monomials, rather than the full set of monomials of degree at most t. This approach combines the border base method from [73] with the generalized flatness condition from [69].

We conclude with illustrating the method on a small example. Consider the polynomial equation: $\mathbf{x}_1^2 + \mathbf{x}_2^2 = 0$, with a unique real root $(0, 0)$ and infinitely many complex roots. Then the moment relaxation of order $t = 1$ has the constraints

$$M_1(y) = \begin{pmatrix} 1 & y_{10} & y_{01} \\ y_{10} & y_{20} & y_{11} \\ y_{01} & y_{11} & y_{02} \end{pmatrix} \succeq 0,\ y_{20} + y_{02} = 0,$$

which imply $y_\alpha = 0$ whenever $\alpha \neq 0$. Therefore the flatness condition holds: $\mathrm{rank} M_1(y) = \mathrm{rank} M_0(y) = 1$. Moreover the kernel of $M_1(y)$ is spanned by the two polynomials $\mathbf{x}_1, \mathbf{x}_2$, which indeed generate the real radical of the ideal $(\mathbf{x}_1^2 + \mathbf{x}_2^2)$.

6. Application to some combinatorial problems

Lift-and-project methods. The polynomial optimization problem (P) contains the general 0/1 linear programming (ILP), asking to optimize a linear function over the $0/1$ solutions to a linear system $Ax \geq b$. Let P denote the integral polytope defined as the convex hull of all $x \in \{0,1\}^n$ satisfying $Ax \geq b$ and let $K = \{x : Ax \geq b\}$ denote its linear relaxation, which can be assumed to lie in the hypercube $[0,1]^n$. A well studied approach in polyhedral combinatorics is to find a (partial) linear inequality description of the polytope P, leading to a new relaxation P' nested between P and K: $P \subseteq P' \subseteq K$, strengthening the initial relaxation K. Several methods have been investigated that construct in a systematic way hierarchies of relaxations nested between P and K, with the property that P is found in finitely many steps. For instance, the classic method in integer programming, which consists of iteratively adding Gomory-Chvátal cuts, finds the integral polytope P in $O(n^2 \log n)$ steps [30], but linear optimization over the first Gomory-Chvátal closure is a hard problem [29]. On the other hand, the lift-and-project methods of Sherali and Adams [96] and of Lovász and Schrijver [71] produce hierarchies of LP and SDP relaxations P_t that find the integral polytope in n steps and with the property that linear optimization over the t-th relaxation P_t is polynomial time for any *fixed* t. They are all based on the following basic strategy:

(a) Generate new polynomial constraints by multiplying the polynomial inequalities $a_j^\mathsf{T} x - b_j \geq 0$ of the system $Ax \geq b$ by x_i or $1 - x_i$ (and their products) and eliminate all squared variables replacing each x_i^2 by x_i.

(b) Linearize all monomials $\prod_{i \in I} x_i$ by introducing new variables y_I, so that the constraints generated in (a) form a linear system in the variables (x, y).

(c) Project back on the x-variables space, which gives a polyhedron P' nested between P and K.

The construction may allow the addition of positive semidefiniteness constraints, leading to stronger semidefinite relaxations. This is the case for the construction of Lovász and Schrijver [71], which we now briefly describe.

Suppose the vector $x \in \{0,1\}^n$ satisfies the system $Ax \geq b$. Consider the new vector $\hat{x} = (1, x) \in \mathbb{R}^{n+1}$ (where the additionnal entry is indexed by '0') and the matrix $Y = \hat{x}\hat{x}^\mathsf{T} \in \mathcal{S}^{n+1}$. Then the matrix Y satisfies the following conditions: (i) $Y \succeq 0$, (ii) $Y_{00} = 1$, (iii) $Y_{0i} = Y_{ii}$ for all $i \in [n]$, and (iv) the vectors $Y^{(i)}$, $Y^{(0)} - Y^{(i)}$ (for $i \in [n]$) satisfy the linear system: $Ax - bx_0 \geq 0$ (where $Y^{(i)} \in \mathbb{R}^{n+1}$ denotes the i-th column of Y). Let $M^+(K)$ denote the set of matrices $Y \in \mathcal{S}^{n+1}$ satisfying the above conditions (i)-(iv), define its projection

$$N^+(K) = \{x \in \mathbb{R}^n : \exists Y \in M^+(K) \text{ such that } x_i = Y_{0i} \ (i \in [n])\},$$

and define analogously $N(K)$ by omitting the positive semidefiniteness condition (i) in the definition of $M^+(K)$. Then, $P \subseteq N^+(K) \subseteq N(K) \subseteq K$. For an integer $t \geq 2$, one can iteratively define $N_t(K) = N(N_{t-1}(K))$, $N_t^+(K) = N^+(N_{t-1}^+(K))$ (setting $N_1(K) = N(K)$ and $N_1^+(K) = N^+(K)$). This leads to hierarchies of linear and semidefinite relaxations, that find P in n steps: $P \subseteq N_t^+(K) \subseteq N_t(K)$, with equality for $t = n$. From the optimization point of view, these hierarchies behave well: if linear optimization over K can be done in polynomial time then the same holds for linear optimization over $N_t(K)$ and $N_t^+(K)$ for any fixed $t \geq 1$ [71].

The paper [71] also investigates in detail how the construction applies to the stable set problem. Given a graph $G = (V = [n], E)$, let $K \subseteq \mathbb{R}^n$ be defined by nonnegativity $x \geq 0$ and the edge inequalities $x_i + x_j \leq 1$ ($\{i,j\} \in E$), so that the corresponding polytope $P = \text{conv}(K \cap \{0,1\}^n)$ is the stable set polytope of G. The first linear relaxation $N(K)$ is completely understood: $N(K)$ is the polyhedron defined by nonnegativity $x \geq 0$ and the odd cycle inequalities $\sum_{i \in O} x_i \leq (|O| - 1)/2$ for all $O \subseteq V$ inducing an odd cycle in G. The relaxation $N^+(K)$ is much stronger. Indeed, for any clique C of G, the corresponding clique inequality $\sum_{i \in C} x_i \leq 1$ is valid for $N^+(K)$, while the first order t for which it is valid for the linear relaxation $N_t(K)$ is $t = |C| - 2$. Moreover the stable set polytope P is found after $\alpha(G)$ steps of the semidefinite hierarchy, compared to $n - \alpha(G) - 1$ steps of the linear hierarchy. These results have motivated much of the interest in these lift-and-project semidefinite relaxations for combinatorial optimization.

The Lasserre approach. The general moment approach applied to (ILP) also produces a hierarchy of semidefinite relaxations $L_t(K)$ converging to P [54]. As explained in [61], the relaxation $L_t(K)$ can easily be described in a direct way following the above lift-and-project strategy. We just indicate here how to apply the previously described general moment method. We start with the set K defined by the polynomial inequalities $g_j = a_j^{\mathsf{T}}\mathbf{x} - b_j \geq 0$ ($j \in [m]$) and the polynomial equations $\mathbf{x}_i^2 - \mathbf{x}_i = 0$ ($i \in [n]$). Then $L_t(K)$ is defined as the set of all vectors $x \in \mathbb{R}^n$ of the form $x = (L(\mathbf{x}_1), \ldots, L(\mathbf{x}_n))$ for some linear functional $L \in \mathbb{R}[\mathbf{x}]_{2t}^*$ satisfying the moment relaxation (MOMt), i.e., the conditions (i) $L(1) = 1$, (ii) $M_t(L) \succeq 0$, (ii) $M_{t-1}(g_j L) \succeq 0$ ($j \in [m]$), and (iii) $L(f) = 0$ for all polynomials f in the truncated ideal $(\mathbf{x}_1^2 - \mathbf{x}_1, \ldots, \mathbf{x}_n^2 - \mathbf{x}_n)_{2t}$.

What the above condition (iii) says is that one can simplify the Lasserre relaxation by eliminating variables and working with smaller moment matrices. Indeed, instead of considering the moment matrix $M_t(L)$ indexed by *all* monomials of degree at most t, it suffices to consider its principal submatrix indexed by all *square-free* monomials of degree at most t (of the form $\prod_{i \in I} x_i$ for $I \in \binom{V}{\leq t}$), and to consider only variables $y_J := L(\prod_{i \in J} x_i)$ for sets $J \in \binom{V}{\leq 2t}$. Here $\binom{V}{\leq t}$ denotes the collection of subsets of $V = [n]$ with cardinality at most t.

As a direct consequence, the flatness condition (4.3) holds at order $t = n + 1$:

$$\text{rank } M_{n+1}(L) = \text{rank } M_n(L).$$

Hence the Lasserre relaxation of order $n + 1$ is exact: $L_{n+1}(K) = P$ (which follows by applying Theorem 4.5). There is also a simple direct proof for this claim or, alternatively, this claim follows from the fact that the Lasserre hierarchy refines the Lovász-Schrijver hierarchy. Namely, for any $t \geq 2$, we have: $L_t(K) \subseteq N(L_{t-1}(K))$, which thus implies the inclusion $L_t(K) \subseteq N_{t-1}(K)$. Moreover, the Lasserre hierarchy also refines the Sherali-Adams hierarchy. We refer to [61] for the above results, and we refer e.g. to the recent work [2] for a comprehensive treatment and further references, also about other lift-and-project hierarchies. We now indicate how the Lasserre hierarchy applies to maximum stable sets, minimum graph colorings and max-cut.

Lasserre hierarchies for $\alpha(G)$ and $\chi(G)$. As an illustration, the moment relaxation (MOMt) for the stable set problem (1.4) reads:

$$\text{las}_t(G) = \max_{y \in \binom{V}{\leq 2t}} \{\sum_{i \in V} y_i : (y_{I \cup J})_{I,J \in \binom{V}{\leq t}} \succeq 0,\ y_{ij} = 0\ (\{i,j\} \in E),\ y_\emptyset = 1\}. \quad (6.1)$$

For $t = 1$, we find Lovász' theta number from (1.6): $\mathrm{las}_1(G) = \vartheta(G)$. Moreover, the Lasserre bound is exact: $\mathrm{las}_t(G) = \alpha(G)$ for $t \geq \alpha(G)$. On the dual side, the sos relaxation (SOSt) asks for the smallest scalar λ for which the polynomial $\lambda - \sum_{i \in V} \mathbf{x}_i$ can be written as a sum of squares of degree at most $2t$ modulo the ideal generated by the polynomials $\mathbf{x}_i\mathbf{x}_j$ (for $\{i,j\} \in E$) and $\mathbf{x}_i^2 - \mathbf{x}_i$ (for $i \in V$). We refer to Gouveia et al. [35] for a detailed study of the hierarchies from this point of view of sums of squares, also in the setting of general polynomial ideals.

In [39] we investigate Lasserre type bounds for the chromatic number $\chi(G)$. A first possibility is to consider the following analogue of the bounds in (6.1):

$$\psi_t(G) = \min_{y \in \binom{V}{\leq 2t}} \{y_\emptyset : (y_{I \cup J})_{I,J \in \binom{V}{\leq t}} \succeq 0,\ y_{ij} = 0\ (\{i,j\} \in E),\ y_i = 1\ (i \in V)\}. \quad (6.2)$$

Then, $\psi_1(G) = \vartheta(\overline{G}) \leq \psi_t(G) \leq \chi(G)$. However, these bounds cannot in general reach the chromatic number since they all remain below the *fractional chromatic number* $\chi_f(G)$: $\psi_t(G) \leq \chi_f(G)$, with equality if $t \geq \alpha(G)$.

To define a hierarchy of semidefinite bounds able to reach the chromatic number $\chi(G)$, one can use the reduction of $\chi(G)$ to the stability number of the cartesian product $G \square K_k$ described in the Introduction. Namely, $\chi(G)$ is equal to the smallest integer k for which $\alpha(G \square K_k) = |V(G)|$. This motivates defining the parameter $\mathrm{Las}_t(G)$ as the smallest integer k for which $\mathrm{las}_t(G \square K_k) = |V(G)|$. Then, we have the inequality: $\mathrm{Las}_t(G) \leq \chi(G)$, with equality for $t = n$. Note that, for $t = 1$, we find again the (rounded) theta number: $\mathrm{Las}_1(G) = \lceil \vartheta(\overline{G}) \rceil$.

An easy way to strengthen the various bounds is by adding the nonnegativity constraint $y \geq 0$ to the program (6.1), call $\mathrm{las}'_t(G)$ the resulting parameter. Analogously, define $\mathrm{Las}'_t(G)$ as the smallest integer k for which $\mathrm{las}'_t(G \square K_k) = |V|$. Then, we have: $\alpha(G) \leq \mathrm{las}'_t(G) \leq \mathrm{las}_t(G)$ and $\mathrm{Las}_t(G) \leq \mathrm{Las}'_t(G) \leq \chi(G)$. It turns out that the parameters $\mathrm{las}'_1(G)$ and $\mathrm{Las}'_1(G)$ coincide, respectively, with the parameters $\vartheta'(G)$ and $\vartheta^+(\overline{G})$ (recall (1.8)).

The bounds $\mathrm{las}_t(G)$ (and $\mathrm{las}'_t(G)$) have been used in particular to upper bound the cardinality of error correcting codes. When dealing with binary codes of length N, one needs to find the stability number of a Hamming graph G, with vertex set $V = \{0,1\}^N$ and where two vertices $u, v \in V$ are adjacent if their Hamming distance does not belong to some prescribed set. Thus this graph G has 2^N vertices. Fortunately it has a large automorphism group which can be used to compute the parameter $\mathrm{las}_t(G)$ with a semidefinite program involving smaller matrices of size $O(N^{2^t-1})$ (polynomial in N for fixed t), while the original formulation (6.1) involves matrices of size $O(|V|^t = 2^{tN})$ (exponential in N). This is shown in [67] using symmetry reduction techniques from [25]. Moreover, Schrijver [93] shows that the semidefinite bound $\mathrm{las}'_1(G) = \vartheta'(G)$ of order $t = 1$ coincides with the well known linear programming bound of Delsarte, which is expressed by a linear program of size N. Furthermore, Schrijver [94] shows that the semidefinite bound of the next order 2 (more precisely, some variation in-between the bounds of order 1 and 2) can be computed with a semidefinite program involving (roughly) $N/2$ matrices of size at most N, which he shows using block-diagonalization techniques for matrix algebras. Numerical computations using these parameters and some strengthenings give the currently best known bounds for codes (see [33, 67, 94] and references therein). Computations for the chromatic number using the bounds $\mathrm{Las}_t(G)$ (and variations) can be found in [39, 41].

The Lasserre hierarchy for max-cut. As another illustration let us apply the Lasserre hierarchy to the max-cut problem (1.2). The equations $\mathbf{x}_i^2 = 1$ permit to express the relaxation (MOMt) as

$$\max_{y \in \mathbb{R}^{\binom{V}{\leq 2t}}} \{ \sum_{\{i,j\} \in E} (w_{ij}/2)(1 - y_{ij}) : (y_{I \Delta J})_{I,J \in \binom{V}{\leq t}} \succeq 0,\ y_\emptyset = 1 \}.$$

For $t = 1$ this is the relaxation (1.3) used by Goemans and Williamson [34] for their 0.878-approximation algorithm for max-cut. More details about geometric properties of the Lasserre hierarchy for max-cut can be found in [63]. A natural question is how many steps are needed to solve max-cut using the hierarchy. In [62] we show that, for the all-ones weight function, the relaxation is exact if and only if $t \geq t_n := \lceil n/2 \rceil$ and we conjecture that t_n iterations suffice for arbitrary weights w. Equivalently, we conjecture that the polynomial $f_w = \mathrm{mc}(G,w) - \sum_{\{i,j\} \in E} (w_{ij}/2)(1 - \mathbf{x}_i \mathbf{x}_j)$ can be written as a sum of squares of degree at most $2t_n$ modulo the ideal $(\mathbf{x}_i^2 - 1 : i \in [n])$. Recently, Blekherman et al. [8] show that this is indeed true when allowing "denominators", i.e., they show that there exists a polynomial p such that $p^2 f_w$ has such a decomposition.

Copositive based hierarchies. Let $\mathcal{C}^n$ denote the copositive cone, consisting of all matrices $M \in \mathcal{S}^n$ for which the polynomial $f_M = \sum_{i,j=1}^n M_{ij} \mathbf{x}_i^2 \mathbf{x}_j^2$ is nonnegative over $\mathbb{R}^n$. As mentioned in the Introduction, the stability number $\alpha(G)$ of a graph G can be obtained from the program (1.9), which is linear optimization over the copositive cone $\mathcal{C}^n$. As we indicate below this formulation leads to another type of hierarchies.

Motivated by the fact that testing matrix copositivity is a hard problem, Parrilo [82] introduced a hierarchy of sufficient conditions, which can be tested using semidefinite optimization and leads to the hierarchy of cones $\mathcal{K}_t$ considered by de Klerk and Pasechnik [24]. Namely, $\mathcal{K}_t$ consists of the matrices $M \in \mathcal{S}^n$ for which the polynomial $f_M (\sum_{i=1}^n \mathbf{x}_i^2)^t$ is a sum of squares. The cone K_0 consists precisely of the matrices M that can be written as the sum of a positive semidefinite matrix and an entrywise nonnegative matrix. Clearly, the cones $\mathcal{K}_t$ form a hierarchy of subcones of $\mathcal{C}^n$: $\mathcal{K}_t \subseteq \mathcal{K}_{t+1} \subseteq \mathcal{C}^n$. Parrilo [82] shows that they cover the interior of $\mathcal{C}^n$: if $f_M(x) > 0$ for all nonzero $x \in \mathbb{R}^n$ then M belongs to some $\mathcal{K}_t$. His proof uses the following result of Pólya: if $g \in \mathbb{R}[\mathbf{x}]$ is a homogeneous polynomial satisfying $g(x) > 0$ for all nonzero $x \in \mathbb{R}^n_+$, then there exists an integer $t \in \mathbb{N}$ for which all the coefficients of the polynomial $(\sum_{i=1}^n \mathbf{x}_i)^t g$ are nonnegative.

The cones $\mathcal{K}_t$ lead to another hierarchy of bounds for the stability number $\alpha(G)$. Starting from relation (1.9), De Klerk and Pasechnik [24] define the parameter

$$\vartheta_t(G) = \min\{\lambda : \lambda(I + A_G) - J \in \mathcal{K}_t\}. \tag{6.3}$$

They show that the first bound is the theta number: $\vartheta_0(G) = \vartheta'(G)$, and they show convergence *after rounding*: $\lfloor \vartheta_t(G) \rfloor = \alpha(G)$ for $t \geq \alpha(G)^2$. Moreover, they conjecture that finite convergence: $\alpha(G) = \vartheta_t(G)$ holds for $t \geq \alpha(G) - 1$, which would mirror the known finite convergence in $\alpha(G)$ steps for the Lasserre bounds $\mathrm{las}_t(G)$. In [38] we give a partial proof and prove this conjecture for all graphs with $\alpha(G) \leq 8$.

This approach also gives lower bounds $\Theta_t(G)$ for the chromatic number $\chi(G)$. Namely, define $\Theta_t(G)$ as the smallest integer k for which $\vartheta_t(G \square K_k) = |V(G)|$. In [38] we compare both types of hierarchies and we show that the Lasserre hierarchies refine these 'copositive based' hierarchies. Namely, we show that $\mathrm{las}'_t(G) \leq \vartheta_{t-1}(G)$ and thus $\Theta_{t-1}(G) \leq$

$\mathrm{Las}'_t(G)$ for any $t \geq 1$. Hence, the Lasserre hierarchy may give better bounds and moreover it seems much easier to handle. For instance its finite convergence is easy, while the finite convergence of the copositive hierarchy is still open. A reason might be that the Lasserre construction uses explicitly the presence of binary variables, while the copositive based construction does not. Nevertheless copositive based approximations have gained popularity in the recent years and they open the way to other types of approaches for approximating hard problems. We refer e.g. to [11, 28] and references therein.

7. Conclusions

We have presented the general approach permitting to construct semidefinite relaxations for polynomial optimization problems by using sums of squares representations for positive polynomials and moment matrices. We reviewed some basic properties regarding in particular their convergence properties. We also discussed how the general methodology applies for building hierarchies of semidefinite relaxations for combinatorial problems in graphs. We have only discussed a small piece of this rapidly expanding research area. We now mention a few other research areas, where this type of methods are also being increasingly used.

Semidefinite optimization and in particular the Lasserre hierarchy are playing a growing role in theoretical computer science for the design of efficient approximation algorithms. Understanding the power and limitations of the Lasserre hierarchy is a fundamental question, which has tight links with complexity theory. For instance, assuming the unique game conjecture [48], Khot et al. [49] show that one cannot beat the Goemans-Williamson 0.878-approximation guarantee for max-cut, which is based on the Lasserre relaxation of smallest order. Yet recent results of Guruswami and Sinop [37] exploit higher order relaxations to give improved approximation algorithms for graph partition problems, depending on spectral properties of the graph. We refer e.g. to [32, 65], the recent overview by Chlemtac and Tulsiani [1, Chap. 6] and references therein.

Semidefinite bounds are also used to attack geometric problems, like the kissing number problem and the problem of coloring the Euclidean space [3, 4]. These problems lead to maximum stable set and minimum coloring problems in infinite graphs. For instance, the kissing number problem is finding a maximum stable set, where the vertex set is the unit sphere with two points being adjacent depending on their spherical distance. Bachoc and Vallentin [3] use low order bounds in the Lasserre hierarchy to give the best known bounds for the kissing number problem, a crucial ingredient in their approach is exploiting symmetry in order to get computable semidefinite programs.

Hierarchies of semidefinite relaxations have also been used recently to attack polynomial optimization problems in noncommutative variables. Such problems arise when, instead of instantiating variables to scalars, one allows variables to be matrices (or bounded operators on some Hilbert space) and they have applications in many areas of quantum phsyics. Given a symmetric polynomial f in n noncommutative variables, one can consider the following two kinds of positivity: f is said to be matrix-positive if $f(X_1, \ldots, X_n) \succeq 0$ when evaluating f at arbitrary matrices $X_1, \ldots, X_n \in \mathcal{S}^d$ ($d \geq 1$), and f is said to be trace-positive if $\mathrm{Tr}(f(X_1, \ldots, X_n)) \geq 0$ for all $X_1, \ldots, X_n \in \mathcal{S}^d$ ($d \geq 1$). These two notions lead to different noncommutative polynomial optimization problems. For both problems analogues of the moment and sums of squares approaches have been investigated, we refer to [12, 20, 84] and references therein.

By Hilbert's theorem, not all nonnegative polynomials are sums of squares. However, Helton [42] shows the following remarkable result: a symmetric polynomial is matrix-positive if and only if it is a sum of Hermitian squares. Moreover, Helton and McCullough [43] show a result characterizing matrix-positivity on a compact set which can be seen as an analogue of Putinar's result (Theorem 3.4). On the other hand, the analogue result for trace-positive polynomials is still open, and it is in fact related to a deep conjecture of Connes [15] in operator algebra. Indeed, Klep and Schweighofer [50] show that Connes' embedding conjecture is equivalent to a real algebraic conjecture characterizing the trace-positive polynomials on all contraction matrices.

Problems in quantum information have led in the recent years to some quantum analogues of the classical graph parameters $\alpha(G)$ and $\chi(G)$. These quantum parameters require to find positive semidefinite matrices satisfying certain polynomial conditions and, as in the classical case, the theta number serves also as bound for them (see [10, 13] and further references therein). Investigating how to construct hierarchies of stronger semidefinite bounds for these quantum graph parameters is a natural direction that we are currently investigating.

References

[1] Anjos, M.A. and Lasserre, J.B. (eds), *Handbook on Semidefinite, Conic and Polynomial Optimization*, International Ser. Oper. Res. & Management Sci. **166** (2012), Springer.

[2] Hu Hin Au and Tuncel, L., *A comprehensive analysis of polyhedral lift-and-project methods*, arXiv:1312.5972.

[3] Bachoc, C. and Vallentin, F., *New upper bounds for kissing numbers from semidefinite programming*, J. Amer. Math. Soc. **21** (2008), 909–924.

[4] Bachoc, C., Nebe, G., de Oliveira Filho, F.M., and Vallentin, F., *Lower bounds for measurable chromatic numbers*, Geom. Funct. Anal. **19** (2009), 645–661.

[5] Ben-Tal, A. and Nemirovski, A., *Lectures on Modern Convex Optimization - Analysis, Algorithms, and Engineering Applications*, MOS-SIAM Series Optim. **2** (2001).

[6] Blekherman, G., *There are significantly more nonnegative polynomials than sums of squares*, Isreal Journal of Mathematics **153** (2006), 355–380.

[7] _____, *Nonnegative polynomials and sums of squares*, Journal of the Amer. Math. Soc. **25** (2012), 617–635.

[8] Blekherman, G., Gouveia, J., Pfeiffer, J., *Sums of squares on the hypercube*, arXiv: 1402.4199, (2014).

[9] Blekherman, G., Thomas, R.R., and Parrilo, P.A. (eds.), *Semidefinite Optimization and Convex Algebraic Geometry*, MOS-SIAM Series Optim. **13** (2012).

[10] Briët, J., Buhrman, H., Laurent, M., Piovesan, T., and Scarpa, G., *Zero-error source-channel coding with entanglement*, arXiv:1308.4283, (2013).

[11] Burer, S., *On the copositive representation of binary and continuous nonconvex quadratic programs*, Math. Program. **120** (2009), 479–495.

[12] Burgdorf, S., *Trace-Positive Polynomials, Sums of Hermitian Squares and The Tracial Moment Problem*, PhD thesis, Universität Konstanz & Université de Rennes 1 (2011).

[13] Cameron, P.J., Montanaro, A., Newman, M.W., Severini, S., and Winter, A., *On the quantum chromatic number of a graph*, Electr. J. Comb. **14**-1(R81) (2007).

[14] Chudnovsky, M., Robertson, N., Seymour, P., and Thomas, R., *The strong perfect graph theorem*, Annals Math. **164**(1) (2006), 51–229.

[15] Connes, A., *Classification of injective factors. Cases* $\Pi_1, \Pi_\infty, \Pi_\lambda$, $\lambda \neq 1$, Ann. of Math. **104**(2) (1976), 73–115.

[16] Cox, D.A., Little, J.B., and O'Shea, D., *Ideals, Varieties and Algorithms*, Springer, 1997.

[17] Curto, R. and Fialkow, L., *Solution of the truncated complex moment problem for flat data*, Memoirs Amer. Math. Soc. **119**(568) (1996).

[18] ———, *Flat extensions of positive moment matrices: recursively generated relations*, Memoirs Amer. Math. Soc. **136**(648) (1998).

[19] ———, *The truncated complex K-moment problem*, Trans. Amer. Math. Soc. **352** (2000), 2825–2855.

[20] Doherty, A.C., Liang, Y.-C., Toner, B., and Wehner, S., *The quantum moment problem and bounds on entangled multi-prover games*, Proc. CCC'08 (2008), 199–210.

[21] De Klerk, E., *Aspects of Semidefinite Programming - Interior Point Algorithms and Selected Applications*, Kluwer, 2002.

[22] De Klerk, E. and Laurent, M., *Error bounds for some semidefinite programming approaches to polynomial minimization on the hypercube*, SIAM J. Optim. **20**(6) (2010), 3104–3120.

[23] ———, *On the Lasserre hierarchy of semidefinite programming relaxations of convex polynomial optimization problems*, SIAM J. Optim. **21** (2011), 824–832.

[24] De Klerk, E. and Pasechnik, D.V., *Approximating the stability number of a graph via copositive programming*, SIAM J. Optim. **12** (2002), 875–892.

[25] De Klerk, E., Pasechnik, D.V., and Schrijver, A., *Reduction of symmetric semidefinite programs using the regular *-representation*, Math. Program. Ser. B **109** (2007), 613–624.

[26] De Oliveira Filho, F., *New Bounds for Geometric Packing and Coloring via Harmonic Analysis and Optimization*, PhD thesis, University of Amsterdam, 2009.

[27] Dickinson, P. and Gijben, L., *On the computational complexity of membership problems for the completely positive cone and its dual*, Comp. Opt. and Appl. **57**(2) (2014), 403–415.

[28] Dür, M., *Copositive programming - a survey*, In Recent Advances in Optimization and its Applications in Engineering, M. Diehl et al. (eds.), Springer, 2010, pp. 3–20.

[29] Eisenbrand, F., *On the membership problem for the elementary closure of a polyhedron*, Combinatorica **19** (2000), 299–300.

[30] Eisenbrand, F. and Schulz, A.S., *Bounds on the Chvátal rank of polytopes in the 0/1 cube*, In G. Cornuéjols et al. IPCO 1999, LNCS **1610** (1999), 137–150.

[31] Garey, M.R. and Johnson, D.S., *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman & Company Publishers (1979).

[32] Gärtner, B. and Matousek, J., *Approximation Algorithms and Semidefinite Programming,* Springer (2012).

[33] Gijswijt, D.C., Mittelmann, H.D., and Schrijver, A., *Semidefinite code bounds based on quadruple distances*, IEEE Trans. Inform. Theory **58** (2012), 2697–2705.

[34] Goemans, M.X. and Williamson, D. *Improved approximation algorithms for maximum cuts and satisfiability problems using semidefinite programming*, J. of the ACM **42** (1995), 1115–1145.

[35] Gouveia, J., Thomas, R.R., and Parrilo, P.A., *Theta bodies for polynomial ideals*, SIAM J. Optim. **20(4)** (2010), 2097–2118.

[36] Grötschel, M., Lovász, L., and Schrijver, A., *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica **1(2)** (1981), 169–197.

[37] Guruswami, V. and Sinop, A., *Lasserre hierarchy, higher eigenvalues, and approximation schemes for quadratic integer programming with PSD objectives*, FOCS 2011, 482–491.

[38] Gvozdenović, N. and Laurent, M., *Semidefinite bounds for the stability number of a graph via sums of squares of polynomials*, Math. Program. Ser. B **110(1)** (2007), 145–173.

[39] _____, *The operator Ψ for the chromatic number of a graph*, SIAM J. Optim. **19(2)** (2008), 572-591.

[40] _____, *Computing semidefinite programming lower bounds for the (fractional) chromatic number via block-diagonalization*, SIAM J. Optim. **19(2)** (2008), 592-615.

[41] Gvozdenović, N., Laurent, M., and Vallentin, F., *Block-diagonal semidefinite programming hierarchies for 0/1 programming*, Oper. Res. Letters **37** (2009), 27-31.

[42] Helton, J.W., *"Positive" noncommutative polynomials are sums of squares*, Ann. of Math. **156** (2002), 675–694.

[43] Helton, J.W. and McCullough, S., *A Positivstellensatz for non-commutative polynopmials*, Trans. Amer. Math. Soc. **356** (2004), 3721–3737.

[44] Henrion, D. and Lasserre, J.B., *Detecting global optimality and extracting solutions in GloptiPoly*, In Positive Polynomials in Control, *LNCIS* **312** (2005), 293–310.

[45] Hilbert, D., *Über die Darstellung definiter Formen als Summe von Formenquadraten*, Math. Annalen **32** (1888), 342–350.

[46] Khachiyan, L. and Porkolab, L., *Computing integral points in convex semi-algebraic sets*, In FOCS (1997), 162–171.

[47] Porkolab, L. and Khachiyan, L., *On the complexity of semidefinite programs*, J. Global Opt. **10** (1997), 351–365.

[48] Khot, S., *On the power of unique 2-prover 1-round games*, In Proc. 34th Ann. ACM Symp. on the Theory of Computing, (2002), 767–775.

[49] Khot, S., Kindler, G., Mossel, E., and O'Donnell, R., *Optimal inapproximability results for MAX-CUT and other 2-variable CSPs?*, In FOCS 2004, 146–154.

[50] Klep, I. and Schweighofer, M., *Connes' embedding conjecture and sums of Hermitian squares*, Adv. Math. **217(4)** (2008), 1816–1837.

[51] ———, *An exact duality theory for semidefinite programming based on sums of squares*, Math. Oper. Res. **38** (2013), 569-590.

[52] Krivine, J.L., *Anneaux préordonnés*, J. Analyse Math. **12** (1964), 307–326.

[53] Lasserre, J.B., *Global optimization with polynomials and the problem of moments*, SIAM J. Optim. **11** (2001), 796–817.

[54] ———, *An explicit exact SDP relaxation for nonlinear* $0-1$ *programs*, In K. Aardal and A.M.H. Gerards (eds.), LNCS **2081** (2001), 293–303.

[55] ———, *A sum of squares approximation of nonnegative polynomials*, SIAM J. Optim. **16** (2006), 751–765.

[56] ———, *Moments, Positive Polynomials and Their Applications*, Imperial College Press (2009).

[57] Lasserre, J.B., Laurent, M., and Rostalski, P., *Semidefinite characterization and computation of real radical ideals*, Foundations Comput. Math. **8** (2008), 607–647.

[58] ———, *A unified approach for real and complex zeros of zero-dimensional ideals*, pages 125–155 in [87].

[59] Lasserrre, J.B., Laurent, M., Mourrain, B., Rostalski, P., and Trebuchet, P., *Moment matrices, border bases and radical computation*, J. Symb. Comput. **51** (2013), 63-85.

[60] Lasserre, J.B. and Netzer, T., *SOS approximations of nonnegative polynomials via simple high degree perturbations*, Math. Zeitschrift **256** (2006), 99–112.

[61] Laurent, M., *A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0-1 programming*, Math. Oper. Res. **28**(3) (2003), 470–496.

[62] ———, *Lower bounds for the number of iterations in semidefinite hierarchies for the cut polytope*, Math. Oper. Res. **28(4)** (2003), 871–883.

[63] ———, *Semidefinite relaxations for Max-Cut*, In The Sharpest Cut, M. Grötschel (ed.), MOS-SIAM Series Optim. **4** (2004), 257–290.

[64] ———, *Revisiting two theorems of Curto and Fialkow on moment matrices*, Proc. Amer. Math. Soc. **133(10)** (2005), 2965–2976.

[65] Laurent, M., Rendl, F., *Semidefinite Programming and Integer Programming*, In Handbook on Discrete Optimization, K. Aardal et al. (eds.), Elsevier (2005), 393-514.

[66] Laurent, M., *Semidefinite representations for finite varieties*, Math. Program. Ser. A **109** (2007), 1–26.

[67] ———, *Strengthened semidefinite programming bounds for codes*, Math. Program. Ser. B **109** (2007), 239–261.

[68] ———, *Sums of squares, moment matrices and optimization over polynomials*, pages 157–270 in [87].

[69] Laurent, M., Mourrain, B., *A generalized flat extension theorem for moment matrices*, Archiv Math. **93(1)** (2009), 87-98.

[70] Lovász, L., *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory **IT-25** (1979), 1–7.

[71] Lovász, L. and Schrijver, A., *Cones of matrices and set-functions and* $0-1$ *optimization*, SIAM J. Optim. **1** (1991), 166–190.

[72] Marshall, M., *Positive Polynomials and Sums of Squares*, Mathematical Surveys and Monographs, Amer. Math. Soc., **146** (2008).

[73] Mourrain, B., *A new criterion for normal form algorithms*, In H. Imai et al. (eds.), LNCS **1719** (1999), 430–443.

[74] Murty, K.G. and Kabadi, S.N., *Some NP-complete problems in quadratic and nonlinear programming*, Math. Program. Ser. A, **39** (1987), 117–129.

[75] Nesterov, Y.E., *Squared functional systems and optimization problems*, In High Performance Optimization, J.B.G. Frenk et al. (eds.), Kluwer (2000), 405–440.

[76] Nesterov, Y.E. and Nemirovski, A., *Interior Point Polynomial Methods in Convex Programming*, SIAM, Studies in Applied Mathematics **13** (1994).

[77] Nie, J., *Certifying convergence of Lasserre's hierarchy via flat truncation*, Math. Program. Ser. A **142** (2013), 485–510.

[78] ———, *Polynomial optimization with real varieties*, SIAM J. Optim. **23(3)** (2013), 1634-1646.

[79] ———, *An exact Jacobian SDP relaxation for polynomial optimization*, Math. Program. Ser. A **137** (2013), 225–255.

[80] ———, *Optimality conditions and finite convergence of Lasserre's hierarchy*, arXiv: 1206.0319v2, (2013).

[81] Nie, J. and Schweighofer, M., *On the complexity of Putinar's Positivstellensatz*, J. Complexity **23(1)** (2007), 135–150.

[82] Parrilo, P.A., *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, PhD thesis, Caltech (2000).

[83] ———, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program. Ser. B **96** (2003), 293–320.

[84] Pironio, S., Navascués, M., and Acín, A., *Convergent relaxations of polynomial optimization problems in non-commutative variables*, SIAM J. Optim. **20(5)** (2010), 2157-2180.

[85] Prestel, A. and Delzell, C.N., *Positive Polynomials - From Hilbert's 17th Problem to Real Algebra*, Springer (2001).

[86] Putinar, M., *Positive polynomials on compact semi-algebraic sets*, Indiana University Math. J. **42** (1993), 969–984.

[87] Putinar, M. and Sullivant, S. (eds.), *Emerging Applications of Algebraic Geometry*, IMA Volumes in Mathematics and its Applications **149** (2009).

[88] Ramana, M.W., *An exact duality theory for semidefinite programming and its complexity implications*, Math. Program. **77** (1997), 129–162.

[89] Reznick, B., *Some concrete aspects of Hilbert's 17th problem*, Contemporary Math. **253** (2000), 251–272.

[90] Scheiderer, C., *Sums of squares of regular functions on real algebraic varieties*, Trans. Amer. Math. Soc. **352** (1999), 1039–1069.

[91] ———, *Positivity and sums of squares: A guide to recent results*, pages 1–54 in [87].

[92] Schmüdgen, K., *The K-moment problem for compact semi-algebraic sets*, Math. Annalen **289** (1991), 203–206.

[93] Schrijver, A., *A comparison of the Delsarte and Lovász bounds*, IEEE Trans. Inform. Theory **25** (1979), 425–429.

[94] ———, *New code upper bounds from the Terwilliger algebra and semidefinite programming*, IEEE Trans. Inform. Theory **51** (2005), 2859–2866.

[95] Schweighofer, M., *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim. **15(3)** (2005), 805–825.

[96] Sherali, H.D. and Adams, W.P., *A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems*, SIAM J. Disc. Math. **3** (1990), 411–430.

[97] Shor, N.Z., *An approach to obtaining global extremums in polynomial mathematical programming problems*, Kibernetika **5** (1987), 102–106.

[98] Stengle, G., *A Nullstellensatz and a Positivstellensatz in semialgebraic geometry*, Math. Ann. **207** (1974), 87–97.

[99] Vandenberghe, L., Boyd, S., *Semidefinite programming*, SIAM Rev. **38**(1996),49–95.

[100] Wolkowicz, H., Saigal, R., and Vandeberghe, L. (eds.), *Handbook of Semidefinite Programming*, Kluwer (2000).

Centrum Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands; Tilburg University, Department of Econometrics and Operations Research, PO Box 90153, 5000 LE Tilburg, The Netherlands.

E-mail: M.Laurent@cwi.nl

Nonsmooth optimization: conditioning, convergence and semi-algebraic models

Adrian S. Lewis

Abstract. Variational analysis has come of age. Long an elegant theoretical toolkit for variational mathematics and nonsmooth optimization, it now increasingly underpins the study of algorithms, and a rich interplay with semi-algebraic geometry illuminates its generic applicability. As an example, alternating projections – a rudimentary but enduring algorithm for exploring the intersection of two arbitrary closed sets – concisely illustrates several far-reaching and interdependent variational ideas. A transversality measure, intuitively an angle and generically nonzero, controls several key properties: the method's linear convergence rate, *a posteriori* error bounds, sensitivity to data perturbations, and robustness relative to problem description. These linked ideas emerge in a wide variety of computational problems. Optimization in particular is rich in examples that depend, around critical points, on "active" manifolds of nearby approximately critical points. Such manifolds, central to classical theoretical and computational optimization, exist generically in the semi-algebraic case. We discuss examples from eigenvalue optimization and stable polynomials in control systems, and a prox-linear algorithm for large-scale composite optimization applications such as machine learning.

Mathematics Subject Classification (2010). Primary 90C31, 49K40, 65K10; Secondary 90C30, 14P10, 93D20.

Keywords. variational analysis, nonsmooth optimization, inverse function, alternating projections, metric regularity, semi-algebraic, convergence rate, condition number, normal cone, transversality, quasi-Newton, eigenvalue optimization, identifiable manifold.

1. Introduction: the Banach fixed point theorem

Our topic — sensitivity and iterative algorithms for numerical inversion and optimization — has deep roots in the Banach fixed point theorem, so we begin our quick introduction there. Given a Euclidean space $\mathbf{E}$ (a finite-dimensional real inner product space), we seek to invert a map $F\colon \mathbf{E} \to \mathbf{E}$. In other words, given a data vector $y \in \mathbf{E}$, we seek a solution vector $x \in \mathbf{E}$ satisfying $F(x) = y$. We analyze this problem around a particular solution $\bar{x} \in \mathbf{E}$ for data $\bar{y} = F(\bar{x})$. A good exposition on the idea of inversion, close in spirit to our approach here, is the monograph of Dontchev and Rockafellar [26].

Given a constant ρ such that the map $I - \rho F$ (where I is the identity) has Lipschitz modulus $\tau = \operatorname{lip}(I - \rho F)(\bar{x}) < 1$ (meaning that the map is locally a strict contraction), Banach's 1922 argument [2] shows that the *Picard iteration*

$$x_{k+1} = x_k - \rho(F(x_k) - y) \quad (\text{for } k = 0, 1, 2, \ldots)$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

converges linearly to a solution $\hat{x}$ of the equation $F(x) = y$, for any data vector y near $\bar{y}$, when initiated near $\bar{x}$. Furthermore, by starting sufficiently near $\bar{x}$, we can ensure an upper bound on the *linear rate* arbitrarily close to τ: in other words, for any constant $\bar{\tau} > \tau$ we know $\bar{\tau}^{-k}|x_k - \hat{x}| \to 0$. This construction shows that the inverse map F^{-1} agrees (graphically) around the point $(\bar{y}, \bar{x})$ with a single-valued function having Lipschitz modulus $\frac{\rho}{1-\tau}$. We thus see the sensitivity of solutions to data perturbations, and error bounds on the distance from approximate solutions x to the true solution in terms of the *a posteriori* error $F(x) - y$.

Similar classical arguments show robustness in the problem description: for linear maps $A\colon \mathbf{E} \to \mathbf{E}$ with norm less than the bound $\frac{\rho}{1-\tau}$, the perturbed map $F + A$ retains (locally) a Lipschitz inverse. Less classically, as we shall see (though related to the Eckart-Young theorem [40]), the bound $\frac{\rho}{1-\tau}$ is optimal.

Consider the even simpler case when the map F is linear, self-adjoint, and positive semidefinite, with maximum and minimum eigenvalues Λ and λ respectively. In the generic case when F is actually positive definite, we could choose $\rho = \frac{1}{\Lambda}$, and then $\tau = 1 - \frac{\lambda}{\Lambda}$. The Picard iteration becomes simply the method of steepest descent for the convex quadratic function $\frac{1}{2}\langle x, Fx\rangle - \langle y, x\rangle$, with constant step size. The key constant $\frac{1}{1-\tau}$ controlling the algorithm's convergence rate, sensitivity, error bounds and robustness, is just $\frac{\Lambda}{\lambda}$, the condition number of F. This constant is also closely associated with the linear convergence rate of other algorithms, such as steepest descent with exact line search, and the method of conjugate gradients.

A broad paradigm, originating with Demmel [22], relates the computational difficulty of a problem instance (here indicated by convergence rate) with the distance to the nearest "ill-posed" instance (in this case one where Lipschitz invertibility breaks down). An extensive theory of Renegar (see [65]), analogous to the theory above for convex quadratic minimization, concerns feasibility and optimization problems with constraints of the form $y \in Fx + K$, for linear maps F and convex cones K: in that case, the algorithms in question are interior-point methods [60].

Over the next couple of sections, we illustrate and study these ideas more broadly. In each case, we consider a computational problem involving inversion or optimization (which amounts to inverting a gradient-type mapping), and study a "regularity" modulus at a particular solution. We observe how that modulus controls error bounds, sensitivity analysis, robustness in the problem description, and the local linear convergence rate of simple iterative algorithms.

2. Variational analysis and alternating projections

We next consider the problem of *set intersection*: given two nonempty closed sets X and Y in the Euclidean space $\mathbf{E}$, we simply seek a point $z \in X \cap Y$. Like our first example, this problem involves a kind of inversion: we seek a point z such that $(0, 0)$ lies in the set $(X - z) \times (Y - z)$, a set we can view as a function of z.

We denote the distance from a point $y \in \mathbf{E}$ to X by $d_X(y)$, and the set of nearest points (or *projection*) by $P_X(y)$. We consider the method of *alternating projections*, which simply repeats the iteration

$$x_{k+1} \in P_X(y_k), \quad y_{k+1} \in P_Y(x_{k+1}).$$

For convex sets, this method has a long history dating back at least to a 1933 work of von Neumann [76], with a well understood convergence theory: a good survey is [3]. While typically slow, its simplicity lends it enduring appeal, even for nonconvex sets. Robust control theory, for example, abounds in low-rank matrix equations, and projecting a matrix M onto the (nonconvex) set of matrices of rank no larger than r is easy: we simply zero out all but the r largest singular values in the singular value decomposition of M (an approach tried in [38], for example). Furthermore, for our current purposes, the method of alternating projections perfectly illustrates many core ideas of variational analysis, as well as our broad thesis.

Central to our discussion is the notion of transversality. If x is a nearest point in the set X to a point $y \in \mathbf{E}$, then any nonnegative multiple of the vector $y - x$ is called a *proximal normal* to X at x: such vectors comprise a cone $N_X^p(x)$. We say that X and Y intersect *transversally* at a point $\bar{z}$ in their intersection when there exists an angle $\theta > 0$ such that the angle between any proximal normal to X and proximal normal to Y, both at points near $\bar{z}$, is always less than $\pi - \theta$. The supremum of such θ is the *transversality angle*. When X and Y are smooth manifolds, transversality generalizes the classical notion [47]. We then have the following special case of a result from [28].

Theorem 2.1 (Convergence of alternating projections). *Initiated near any transversal intersection point for two closed sets, the method of alternating projections converges linearly to a point in the intersection. If the transversality angle is $\bar{\theta} > 0$, then we can ensure an upper bound on the convergence rate arbitrarily close to $\cos^2(\frac{\bar{\theta}}{2})$ by initiating sufficiently near the intersection point.*

Notable in this result (unlike all previous analysis, such as [52]) is the absence of any assumptions on the two intersecting sets, such as convexity or smoothness. Central to the proof is the Ekeland variational principle [35].

Modern variational analysis grew out of attempts to expand the broad success of convex analysis — an area for which Rockafellar's seminal monograph [67] remains canonical — and to unify it with classical smooth analysis. Classical analysis relies crucially on limiting constructions: for example, the definition of transversally intersecting smooth manifolds (a special case of our property) involves their tangent spaces. The more general property described above also has a limiting flavor, and we can express it more succinctly using a limiting construction. This construction originated in Clarke's 1973 thesis [16, 17], in a convexified form, and a couple of years later, in the raw form we describe here (including implications for transversality) in work reported in Mordukhovich's paper [57] along with contemporaneous joint studies with Kruger ranging from [59] to [45]. It is fundamental to variational analysis: the expository monographs [8, 18, 58, 68] each provide excellent surveys and historical discussion, [7] is a gentler introduction, and [43, p. 112] recounts some early history. The monograph [26] is particularly attuned to our approach here.

A limit of proximal normals to the set X at a sequence of points approaching a point $x \in X$ is simply called a *normal* at x. Such vectors comprise a closed cone $N_X(x)$, possibly nonconvex, called the *normal cone*. With this notation, transversality at the point $\bar{z}$ is simply the property

$$N_X(\bar{z}) \cap -N_Y(\bar{z}) = \{0\},$$

and the transversality angle is the minimal angle between pairs of vectors in the cones $N_X(\bar{z})$ and $-N_Y(\bar{z})$.

The idea of a normal vector to a closed set $X \subset \mathbf{E}$ is a special case of the idea of a "subgradient" of a lower semicontinuous extended-real-valued function on $\mathbf{E}$. For simplicity of exposition, in this essay we confine ourselves to properties of normals, but many of the results that we present extend to subgradients.

The terminology of "normals" we use here is consistent with classical usage for smooth manifolds and convex sets, a fact fruitfully seen in a broader context. A set $X \subset \mathbf{E}$ is nonempty, closed and convex if and only if its projection mapping P_X is everywhere single-valued. More generally [64], X is *prox-regular* at a point $x \in X$ when P_X is everywhere single-valued nearby. In that case, the limiting construction above is superfluous: all normals are proximal, so the cones $N_X(x)$ and $N_X^p(x)$ coincide (and are closed and convex). Prox-regularity applies more broadly than convexity, to smooth manifolds, for example. A set $\mathcal{M} \subset \mathbf{E}$ is a $\mathcal{C}^{(2)}$ *manifold* around a point $\bar{x} \in \mathcal{M}$ if it can be described locally as $F^{-1}(0)$, where the map $F\colon \mathbf{E} \to \mathbf{F}$ is twice continuously differentiable, with surjective derivative at $\bar{x}$. In that case, classical analysis shows $\mathcal{M}$ is prox-regular at $\bar{x}$.

For convex sets X and Y, transversality fails at a common point exactly when there exists a separating hyperplane through that point; a small translation of one set then destroys the intersection. The following result [45], a local generalization of the separating hyperplane theorem, hints at the power of transversality.

Theorem 2.2 (Extremal principle). *On any neighborhood of a point where two closed sets intersect transversally, all small translations of the sets must intersect.*

This principle is a unifying theme in the exposition [58], for example. One proof proceeds constructively, using alternating projections [52].

Another consequence of transversality is the existence of an error bound, discussed in [41, p. 548], estimating the distance to the intersection of the two sets in terms of the distances to each separately. Notice, in the product space $\mathbf{E} \times \mathbf{E}$, we have the relationship $d_{X\times Y}(z,z) = \sqrt{d_X^2(z) + d_Y^2(z)}$ for any point $z \in \mathbf{E}$.

Theorem 2.3 (Error bound). *If closed sets X and Y intersect transversally at a point $\bar{z}$, then there exists a constant $\rho > 0$ such that all points z near $\bar{z}$ satisfy*

$$d_{X\cap Y}(z) \leq \rho d_{X\times Y}(z,z).$$

Intuitively, when the transversality angle is small, we expect to need a large constant ρ in the error bound above. We can make this precise through a single result, discussed in [52], subsuming the preceding two.

Theorem 2.4 (Sensitivity). *Sets X and Y intersect transversally at a point $\bar{z}$ if and only if there exists a constant $\rho > 0$ such that all points z near $\bar{z}$ and all small translations X' of X and Y' of Y satisfy*

$$d_{X'\cap Y'}(z) \leq \rho d_{X'\times Y'}(z,z).$$

The infimum of such ρ is $(1-\cos\bar{\theta})^{-\frac{1}{2}}$, where $\bar{\theta}$ is the transversality angle.

We see a pattern of ideas analogous to those for inversion via the Picard iteration: an algorithm whose linear convergence rate is governed by a sensitivity modulus. To pursue the analogy intuitively a little further, when the transversality angle $\bar{\theta}$ is small, we expect a small change in the problem description to destroy transversality. To illustrate, suppose $\bar{z} = 0$, and at that point choose unit normals u and v to the sets X and Y respectively with an angle

of $\pi - \bar{\theta}$ between them. Now consider the orthogonal map R on the space $\mathbf{E}$ rotating the u-v plane through an angle $\bar{\theta}$ and leaving its orthogonal complement invariant, and so that $Ru = -v$. The sets RX and Y are no longer transversal at zero, since Ru is normal to RX. A natural way to measure the change in the problem description is to view the original problem as $(I, I)z \in X \times Y$, and the perturbed problem as $(R^{-1}, I)z \in X \times Y$, the size of the change being the norm $\|I - R^{-1}\| = 2 \sin \frac{\bar{\theta}}{2}$. In Section 4, where we consider the broader pattern, we see that in fact a somewhat smaller change will destroy transversality. However, rather than pursue the analogy further now, we first consider whether transversality is a realistic assumption in concrete settings.

3. Generic transversality of semi-algebraic sets

Like most areas of analysis, the reach of general variational analysis is limited by pathological examples. In our present context, for example, consider the set intersection problem in $\mathbf{R}^3 = \mathbf{R}^2 \times \mathbf{R}$, for the two sets $X = \mathbf{R}^2 \times \{0\}$ and

$$Y \;=\; \big\{(w, r) : r \geq f(w)\big\},$$

where the function f is a famous 1935 example of Whitney [77] that is continuously differentiable and has an arc of critical points with values ranging from -1 to 1. Thus for every number s in the interval $[-1, 1]$ there exists a critical point w with $f(w) = s$: hence the vector $(0, -1)$ is normal to Y at the point (w, s), so clearly the intersection of the translated set $X + (x, s)$ (for any point $x \in \mathbf{R}^2$) and the set Y is not transversal at the point (w, s). We have arrived at an example of two closed sets for which, after translations, the failure of transversality is not uncommon.

On the other hand, in concrete computational settings we do not expect to encounter Whitney's example. To be more precise, we take as an illustrative model of "concrete" computation the world of *semi-algebraic* sets. We view the Euclidean space $\mathbf{E}$ as isomorphic to the space $\mathbf{R}^n$ (for some dimension n), and consider finite unions of sets, each defined by finitely-many polynomial inequalities. This world, and its generalizations in models of "tame" geometry first promoted by Grothendieck [39], strike happy compromises between broad generality and good behavior. Concise and clear surveys appear in [19, 20, 74].

On the one hand, semi-algebraic sets comprise a rich class: in particular, they may be neither convex nor smooth. They are, furthermore, often easy to recognize without recourse to the basic definition, due to the Tarski-Seidenberg Theorem: the projection of a semi-algebraic set onto a subspace is semi-algebraic. Applying this principle repeatedly shows that sets like the cone of real positive semidefinite symmetric matrices and sets of matrices of bounded rank are semi-algebraic.

On the other hand, semi-algebraic sets cannot be too pathological (or "wild", in Grothendieck's terminology). For example, although nonsmooth in general, they stratify into finite unions of analytic manifolds, so have a natural notion of *dimension*, namely the largest dimension of any manifold in a stratification. Another important example for us concerns the term "generic". In this essay, we call a property that depends on a data vector y in a Euclidean space $\mathbf{E}$ *generic* when it holds except for y in a set $Z \subset \mathbf{E}$ of measure zero. Unlike the general case, if Z is semi-algebraic, then the following properties are equivalent:

- Z has measure zero.

- Z has dimension strictly less than that of $\mathbf{E}$.
- the complement of Z is dense.
- the complement of Z is topologically generic.

We call semi-algebraic sets Z with these properties *negligible*.

No semi-algebraic analog can exist of the example we constructed from Whitney's function. Specifically, we have the following result [28], a special case of a powerful generalization we discuss later.

Theorem 3.1 (Generic transversality). *Suppose X and Y are semi-algebraic subsets of $\mathbf{E}$. Then for all vectors z outside a negligible semi-algebraic subset of $\mathbf{E}$, transversality holds at every point in the intersection of the sets $X - z$ and Y.*

Practical variational problems are often highly structured, involving sparse data, for example. Nonetheless, this result is reassuring: it suggests that, for concrete intersection problems with sets subject to unstructured perturbations, transversality is a reasonable assumption.

4. Measuring invertibility: metric regularity

Our sketch hints at an intriguing web of ideas concerning computational inversion:

- *Sensitivity* of solutions to data perturbation
- *Linear error bounds* for trial solutions in terms of measured error
- *Robustness* in problem description
- Local *linear convergence* of simple solution algorithms.

A single *modulus* (a condition number or angle in our examples) quantifies all four properties. We call problem instances *well-posed* when the modulus is finite, and, within broad problem classes, this property is *generic*. As we now describe, these interdependent ideas are very pervasive indeed.

To capture the abstract idea of computational inversion, we consider two Euclidean spaces $\mathbf{E}$ and $\mathbf{F}$ and a set-valued mapping Φ on $\mathbf{E}$ whose images are subsets of $\mathbf{F}$: we write $\Phi\colon \mathbf{E} \rightrightarrows \mathbf{F}$. Given a data vector $\bar{y} \in \mathbf{F}$, our problem is to find a solution $x \in \mathbf{E}$ to the generalized equation $\bar{y} \in \Phi(x)$. This model subsumes, of course, the example of a classical equation, when Φ is single-valued and smooth, but it is much more versatile than its abstract simplicity might suggest, modeling inequalities rather than just equations, for instance.

To illustrate the power of the approach, among many further examples, we keep in mind two in particular. The first we have seen already. Given two sets X and Y in the space $\mathbf{E}$, if we consider the mapping

$$\Phi\colon \mathbf{E} \rightrightarrows \mathbf{E}^2 \text{ defined by } \Phi(z) = (X - z) \times (Y - z), \tag{4.1}$$

then the problem $0 \in \Phi(z)$ is just set intersection.

For the second example, we return to the normal cone $N_X(x)$ to a nonempty closed set X in $\mathbf{E}$, but now thought of as a mapping $N_X\colon \mathbf{E} \rightrightarrows \mathbf{E}$ (defining $N_X(x) = \emptyset$ for $x \notin \mathbf{E}$). Solutions $x \in \mathbf{E}$ of the generalized equation $\bar{y} \in N_X(x)$ are *critical points* for the linear

optimization problem $\sup_X \langle \bar{y}, \cdot \rangle$. This terminology is in keeping with the classical notion when X is a smooth manifold, while for convex X, critical points are just maximizers. For simplicity, this essay concentrates on linear rather than general optimization. However, that restriction involves little loss of generality: for example, minimizing a function $f\colon \mathbf{E} \to \mathbf{R}$ is equivalent to a linear optimization problem over the *epigraph* of f:

$$\inf\{\tau : (x,\tau) \in \operatorname{epi} f\}, \quad \text{where} \quad \operatorname{epi} f \;=\; \big\{(x,\tau) \in \mathbf{E} \times \mathbf{R} : \tau \geq f(x)\big\}.$$

The fundamental idea, unifying the kinds of error bounds and sensitivity analysis we have illustrated so far, is *metric regularity* of the mapping Φ at a point $\bar{x} \in \mathbf{E}$ for a data vector $\bar{y} \in \Phi(\bar{x})$: the existence of a constant $\rho > 0$ such that

$$d_{\Phi^{-1}(y)}(x) \leq \rho d_{\Phi(x)}(y) \;\text{ for all } (x,y) \text{ near } (\bar{x},\bar{y}). \tag{4.2}$$

We call $\bar{y}$ a *critical value* if Φ is not metrically regular for $\bar{y}$ at some point in $\mathbf{E}$.

Inequality (4.2) is a locally uniform linear bound on the error between a trial solution x and the true solution set $\Phi^{-1}(y)$ for data y, in terms of the measured error from y to the trial image $\Phi(x)$. It captures both error bounds (where $y = \bar{y}$) and sensitivity analysis (where y varies). In highlighting metric regularity, we are implicitly supposing inversion to be computationally hard: the set $\Phi(x)$ is more tractable than the set $\Phi^{-1}(y)$.

The mapping Φ is *closed* when its graph

$$\operatorname{gph} \Phi \;=\; \big\{(x,y) \in \mathbf{E} \times \mathbf{F} : y \in \Phi(x)\big\}$$

is closed. It is *semi-algebraic* when its graph is semi-algebraic, and then its *graphical dimension* is the dimension of its graph. Around any point $(\bar{x},\bar{y}) \in \operatorname{gph} \Phi$ we define three constants:

- The *modulus* is the infimum of the constants $\rho > 0$ such that the metric regularity inequality (4.2) holds.
- The *radius* is the infimum of the norms of linear maps $G\colon \mathbf{E} \to \mathbf{F}$ such that the mapping $\Phi + G$ is not metrically regular at $\bar{x}$ for $\bar{y} + G\bar{x}$.
- The *angle* is the transversality angle for the sets $\operatorname{gph} \Phi$ and $\mathbf{E} \times \{\bar{y}\}$ at the point $(\bar{x},\bar{y})$.

These quantities are strongly reminiscent of the linked ideas opening this section. The first constant quantifies error bounds and sensitivity. The second concerns how robust the problem is under linear perturbations: within that class, it measures the distance to the nearest ill-posed (metrically irregular) instance. By Theorem 2.1, the third quantity controls the local linear convergence rate of at least one simple conceptual algorithm for finding a solution x near $\bar{x}$ to the generalized equation $\bar{y} \in \Phi(x)$: alternating projections on the sets $\operatorname{gph} \Phi$ and $\mathbf{E} \times \{\bar{y}\}$.

In this essay we concentrate on what we loosely call "simple" algorithms, relying only on basic evaluations and properties of the mapping Φ. By contrast, Newton-type schemes use, or assume and approximate, tangential ("higher-order") properties of $\operatorname{gph} \Phi$. For an extensive discussion relating metric regularity and the convergence of Newton-type methods, see [26]. The conceptual algorithm above belongs to the class of *proximal point* methods, which minimize functions f using the iteration

$$x_{k+1} \;\in\; \operatorname{argmin}\big\{f(x) + |x - x_k|^2\big\}.$$

In this case, $f(x) = d^2_{\Phi(x)}(\bar{y})$.

Between the three diverse quantities we have introduced, we have the following extraordinarily simple and general relationship.

Theorem 4.1 (Metric regularity). *At any point in the graph of any closed set-valued mapping we have*

$$\text{radius} \;=\; \frac{1}{\text{modulus}} \;=\; \tan(\text{angle}).$$

The first equality is [25, Theorem 1.5], while the second is a version of the "coderivative criterion" for metric regularity (whose history is discussed in [68, p. 418]). For reasons of space, we omit a dual "derivative criterion", expressible using tangents in the place of normals: see [24] for a discussion.

For instance, consider our motivating example, the intersection problem for two sets X and Y discussed in Section 2. Equation (4.1) describes the corresponding mapping. Theorem 2.4 (Sensitivity) and a calculation [52] shows that the modulus is $(1 - \cos\bar{\theta})^{-\frac{1}{2}}$, where $\bar{\theta}$ is the transversality angle for X and Y at the intersection point. The radius is therefore $\sqrt{2}\sin\frac{\bar{\theta}}{2}$. The proximal point method above is that for minimizing the function $d^2_X + d^2_Y$.

We consider our earlier example of normal cone operators shortly. First, however, we return to the classical single-valued case. When the mapping Φ is linear, standard linear algebra shows that metric regularity is equivalent to surjectivity, and the Eckart-Young Theorem identifies the radius as just the smallest singular value of Φ. More generally, for continuously differentiable Φ, the Lusternik-Graves Theorem amounts to the fact that the modulus of Φ at any point $\bar{x}$ agrees with that of its linear approximation there (see [26]), and hence equals the reciprocal of the smallest singular value of the derivative of Φ at $\bar{x}$.

This classical case also guides us on the question of whether metric regularity is a generic property. In this case, the set of critical values $C \subset \mathbf{F}$ is just the image under the mapping Φ of the set in $\mathbf{E}$ where the derivative of Φ is not surjective. The example of Whitney that we discussed earlier shows that C may be large (in that case an interval in $\mathbf{R}$) even for continuously differentiable Φ. However, assuming Φ is sufficiently smooth, Sard's theorem [69] guarantees that C has measure zero. In this sense, metric regularity is typical.

To address this question for set-valued mappings, we consider the semi-algebraic world, in which we have the following striking result of Ioffe [42].

Theorem 4.2 (Semi-algebraic Sard). *The set of critical values of any semi-algebraic set-valued mapping is semi-algebraic and negligible.*

In computational practice, generic results like this one may often be of limited consequence, since generalized equations often involve highly structured data. Nonetheless, like its special case, Theorem 3.1 (Generic transversality), the result provides a reassuring baseline: for concrete generalized equations with unstructured data, metric regularity is a reasonable assumption.

5. Interlude: nonsmooth optimization via quasi-Newton methods

Metric regularity spans a broad range of inversion and optimization problems. Its suggestive links to convergence rates tempt us to study linearly convergent algorithms, whenever we encounter them, through the lens of metric regularity. An important recurrent theme in the

work of the late Paul Tseng, for example, was the use of error bounds in linear convergence results [56].

An intriguing case is the popular BFGS method [61] (named for its inventors, Broyden, Fletcher, Goldfarb and Shanno) for minimizing a function $f\colon \mathbf{R}^n \to \mathbf{R}$. The BFGS algorithm is a quasi-Newton method, so called by association with the Newton iteration for minimizing a $\mathcal{C}^{(2)}$ function f:

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k).$$

The BFGS method replaces the inverse Hessian by an approximation H_k in the space of n-by-n symmetric matrices $\mathbf{S}^n$, and involves a step length $\alpha_k > 0$:

$$x_{k+1} = x_k - \alpha_k H_k \nabla f(x_k).$$

We then choose H_{k+1} to be the minimizer over the positive-definite matrices of the strictly convex function

$$H \;\mapsto\; \text{trace}(H_k^{-1} H) - \ln\det H \tag{5.1}$$

(see [36]), subject to a linear constraint called the *secant condition*:

$$H\big(\nabla f(x_{k+1}) - \nabla f(x_k)\big) \;=\; x_{k+1} - x_k.$$

The secant condition forces H_{k+1} to behave like the true inverse Hessian in the direction of the last step taken, while the objective (5.1) keeps H_{k+1} close to H_k, since its unconstrained minimizer is at H_k. A simple formula [61] expresses H_{k+1} explicitly as a rank-two perturbation of H_k.

The step length α_k is chosen by a line search on the univariate function

$$\alpha \;\mapsto\; h(\alpha) \;=\; f\big(x_k - \alpha H_k \nabla f(x_k)\big),$$

aiming to satisfy two conditions (called the Armijo and Wolfe conditions):

$$h(\alpha) < h(0) + c_1 h'(0)\alpha \;\text{ and }\; h'(\alpha) > c_2 h'(0).$$

The constants $c_1 < c_2$ in the interval $(0, 1)$ are fixed at the outset. The Armijo condition requires the decrease in the value of h to be a reasonable fraction of its instantaneous decrease at zero, while the Wolfe condition prohibits steps that are too small, by requiring a reasonable reduction in the rate of decrease in h, and ensures the existence of a positive-definite H_{k+1} satisfying the secant condition. A simple bisection scheme finds a suitable step α_k by maintaining the endpoints of a search interval such that the Armijo condition holds on the left and fails on the right, checking both conditions at the midpoint, and then halving the interval accordingly. For a thorough description, see [53].

The BFGS algorithm has been a method of choice for smooth minimization for several decades. It is robust and fast, typically converging superlinearly to a local minimizer. Given its motivation — approximating a Hessian — it seems astonishing that the algorithm also serves as an excellent general-purpose method for *nonsmooth* nonconvex minimization. In principle the algorithm might encounter a point x_k at which the function f is not differentiable, and thereby break down, but with generic initialization, no such breakdowns seem to occur.

A systematic study [53] investigated this phenomenon, and we return to an example from that study later in this work. In general, the BFGS method, when applied to minimize a

semi-algebraic Lipschitz function and with generic initialization, always seems to generate a sequence of function values (including all those computed in the line search) that converges to a stationary value — the value of the function at a point near which convex combinations of gradients are arbitrarily small. Furthermore, for nonsmooth stationary values, that convergence is *linear*! Our current context demands the obvious question: does some condition number or modulus of metric regularity govern the convergence rate of the BFGS method on nonsmooth problems? We seem far from any understanding of this question.

6. Strong regularity and second-order properties

One way to strengthen the metric regularity property is especially important for sensitivity analysis and numerical methods. A set-valued mapping $\Phi\colon \mathbf{E} \rightrightarrows \mathbf{F}$ is clearly metrically regular at a point $\bar{x}$ for a value $\bar{y}$ when the graph of the inverse mapping Φ^{-1} coincides locally with the graph of a single-valued Lipschitz map $G\colon \mathbf{F} \to \mathbf{E}$ around the point $(\bar{y}, \bar{x})$. In that case, we call Φ *strongly metrically regular* (terminology deriving from Robinson [66]); the regularity modulus coincides with the Lipschitz modulus $\operatorname{lip} G(\bar{y})$. We call $\bar{y}$ a *weakly critical value* if there exists a point in $\mathbf{E}$ at which Φ is not strongly metrically regular for $\bar{y}$.

We began, in Section 1, with an example of strong metric regularity: a single-valued map $\Phi\colon \mathbf{E} \to \mathbf{E}$ such that $I - \rho\Phi$ is, locally, a strict contraction. A related example derives from the setting of the classical inverse function theorem: a continuously differentiable map $\Phi\colon \mathbf{E} \to \mathbf{E}$ is strongly metrically regular at points where the derivative of Φ is invertible.

Less classically, suppose the set $\mathcal{M} \subset \mathbf{E}$ is a $\mathcal{C}^{(2)}$ manifold around a point $\bar{x} \in \mathcal{M}$, and consider the mapping Φ defined in terms of the normal cone $N_{\mathcal{M}}$ by

$$\Phi(x) = \begin{cases} x + N_{\mathcal{M}}(x) & (x \in \mathcal{M}) \\ \emptyset & (x \notin \mathcal{M}). \end{cases}$$

Strong metric regularity holds at $\bar{x}$ for $\bar{x}$, because around the point $(\bar{x}, \bar{x})$, the inverse mapping Φ^{-1} agrees graphically with the projection operator $P_{\mathcal{M}}$, which is single-valued and Lipschitz.

The previous section included a conceptual algorithm for solving metrically regular generalized equations, whose convergence rate is controlled by the modulus. Assuming, instead, strong metric regularity (of the mapping Φ at the point $\bar{x}$, for the value 0, say), Pennanen [62] linked the modulus to the convergence rates of algorithms closer to computational practice (in "multiplier methods"). For example, for any constant c larger than twice the modulus, there exists a neighborhood U of $\bar{x}$ such that, with initial point $x_0 \in U$, the proximal-point-type iteration

$$x_k - x_{k+1} \in c\Phi(x_{k+1}), \quad \text{with } x_{k+1} \in U \tag{6.1}$$

always generates sequences converging linearly to a solution of the generalized equation $0 \in \Phi(x)$. (The bound on the rate behaves, as $c \to \infty$, like $\frac{\sqrt{5}}{c}$ times the modulus.) In keeping with our focus on simple algorithms, we pass by the strong connections between strong metric regularity and Newton methods (most importantly in numerical optimization via sequential quadratic programming). That line of investigation, first pursued in [44], is discussed at length in the monograph [26].

Unlike critical values, weakly critical values may be common, even for semi-algebraic mappings. For example, mapping every point to the whole range space $\mathbf{F}$ results in every value being weakly critical. As the next result [32] makes clear, however, this behavior can only result from a large graph.

Theorem 6.1 (Strong semi-algebraic Sard). *If a set-valued mapping $\Phi\colon \mathbf{E} \rightrightarrows \mathbf{F}$ is semi-algebraic and has graphical dimension no larger than $\dim \mathbf{F}$, then its set of weakly critical values is semi-algebraic and negligible.*

This result applies to single-valued semi-algebraic maps $\Phi\colon \mathbf{E} \to \mathbf{E}$ in particular. More interesting for optimizers, however, is the following corollary [27, 30].

Theorem 6.2 (Normal cone mapping). *For any closed semi-algebraic set $X \subset \mathbf{E}$, the normal cone mapping $N_X\colon \mathbf{E} \rightrightarrows \mathbf{E}$ has graphical dimension equal to $\dim \mathbf{E}$, and hence its set of weakly critical values is semi-algebraic and negligible.*

This result suggest that, for concrete linear optimization problems over a set $X \subset \mathbf{E}$ with unstructured objective $\langle \bar{y}, \cdot \rangle$ and solution $\bar{x}$, strong metric regularity of the normal cone mapping N_X is a reasonable assumption. As the next result [31] reveals, this type of property is closely related to second-order conditions in optimization, classically guaranteeing quadratic growth via a Hessian condition. Following our pared-down approach, we focus on linear optimization, but, as before, we could consider a more general problem of the form $\inf_X f$ as seeking a point (x, τ) in the set epi $f \cap (X \times \mathbf{R})$ to maximize the linear function $-\tau$.

Theorem 6.3 (Strong regularity and quadratic growth). *Given a closed set X and a vector $\bar{y} \in \mathbf{E}$, suppose that the point $\bar{x} \in X$ is a local maximizer of the linear function $\langle \bar{y}, \cdot \rangle$ over X. Consider the following three properties:*

- *The normal cone mapping N_X is strongly metrically regular at $\bar{x}$ for $\bar{y}$.*
- *For some scalar $\kappa > 0$ and neighborhood U of $\bar{x}$, "uniform quadratic growth" holds: for all vectors y near $\bar{y}$, there exists a point $x \in X \cap U$ so*

$$\langle y, x' \rangle \leq \langle y, x \rangle - \kappa |x' - x|^2 \ \text{ for all } \ x' \in X \cap U. \tag{6.2}$$

- *The "negative definite" condition holds:*

$$(z, w) \in N_{\operatorname{gph} N_X}(\bar{x}, \bar{y}) \ \text{ and } \ w \neq 0 \implies \langle z, w \rangle < 0.$$

In general, the first condition implies the second two, and so, if X is semi-algebraic, then for all $\bar{y}$ outside a negligible semi-algebraic set, all three conditions hold. If, on the other hand, X is prox-regular at $\bar{x}$, then all three conditions are equivalent.

This result has multiple roots, and deserves some comments. Bonnans and Shapiro include a careful study of uniform quadratic growth in their monograph [6]. Geometrically, condition (6.2) describes a ball with surface containing the point x, center on the ray $x - \mathbf{R}_+ y$, and containing the set X around x. It is natural to include the final special case of a prox-regular set, because the first condition alone turns out to imply a local form of prox-regularity [31]. Assuming prox-regularity, a fourth equivalent notion is *tilt stability* [63].

The link with second-order conditions is not surprising, because the regularity modulus of the normal cone mapping is related via Theorem 4.1 (Metric regularity) to the transversality angle at the intersection point $(\bar{x}, \bar{y})$ for the sets $\operatorname{gph} N_X$ and $\mathbf{E} \times \{\bar{y}\}$, which in turn is just the minimal angle between the subspace $\{0\} \times \mathbf{F}$ and the cone

$$N_{\operatorname{gph} N_X}(\bar{x}, \bar{y})$$

appearing in the third condition. Mordukhovich [58] uses exactly this iterated normal cone construction to define his *generalized Hessian*.

For a semi-algebraic set X, the result above, while interesting, dramatically understates the good behavior of N_X^{-1}: it will typically be single-valued and not just Lipschitz but *analytic*. We explore far-reaching consequences next.

7. Identifiability and the active set philosophy

Given a closed set $X \subset \mathbf{E}$ and a data vector $\bar{y} \in \mathbf{E}$, consider once again the linear optimization problem

$$\sup_X \langle \bar{y}, \cdot \rangle. \tag{7.1}$$

Recall that a point $x \in X$ is *critical* when $\bar{y} \in N_X(x)$.

A wide variety of iterative methods for the linear optimization problem generate *asymptotically critical* sequences (x_k) in X for $\bar{y}$, meaning that some sequence of normals $y_k \in N_X(x_k)$ converges to $\bar{y}$ (implying in particular that any limit point of (x_k) is critical for the problem (7.1)). We aim to profit from this behavior by simplifying the possibly complicated underlying set X. We first illustrate with two examples: alternating projections, and proximal point methods.

Suppose we seek a critical point for our linear optimization problem by applying the proximal point method (6.1) to the mapping defined on $\mathbf{E}$ by $x \mapsto \Phi(x) = N_X(x) - \bar{y}$. Assume the normal cone mapping N_X is strongly metrically regular at $\bar{x}$ for $\bar{y}$. We arrive at the following relationship between iterates, in a neighborhood of a solution $\bar{x}$:

$$\frac{1}{c}(x_k - x_{k+1}) + \bar{y} \in N_X(x_{k+1}).$$

This uniquely defines a sequence in a neighborhood of any fixed solution that converges, providing the constant c is large enough. Since the left-hand side must therefore converge to $\bar{y}$, the sequence of iterates (x_k) is asymptotically critical.

As another example, given two closed sets X and Y in $\mathbf{E}$, we could rewrite the set intersection problem as the linear optimization problem $\sup\{-\tau : (x, y, \tau) \in S\}$, where $S \subset \mathbf{E}^2 \times \mathbf{R}$ is the set defined by the constraint $\tau \geq \frac{1}{2}|x - y|^2$. A quick calculation shows that the method of alternating projections generates two sequences of points, $x_k \in X$ and $y_k \in Y$, satisfying

$$(0, x_k - x_{k+1}, -1) \in N_S(s_k), \quad \text{where } s_k = \Big(x_{k+1}, y_k, \frac{1}{2}|x_{k+1} - y_k|^2\Big).$$

Under reasonable conditions — those of Theorem 2.1 (Convergence of alternating projections), for example — we know that both sequences converge to a point $z \in X \cap Y$. Hence the sequence (s_k) is asymptotically critical for the problem.

We now introduce a simple but powerful variational idea for the set X. We call a subset $\mathcal{M} \subset X$ *identifiable* at a point $\bar{x} \in X$ for the vector $\bar{y} \in \mathbf{E}$ if every asymptotically critical sequence for $\bar{y}$ converging to $\bar{x}$ must eventually lie in $\mathcal{M}$. If $\mathcal{M}$ is also a $\mathcal{C}^{(2)}$ manifold at $\bar{x}$, then we simply call it an *identifiable manifold* at $\bar{x}$ for $\bar{y}$. Such a subset hence balances two competing demands: as a subset of the typically nonsmooth set X, it must be small enough to be a smooth manifold, and yet large enough to capture the tail of every asymptotically critical sequence.

The existence of an identifiable manifold seems, at first sight, a demanding condition. The next result [29], on the other hand, shows that such a manifold uniquely captures important sensitivity information about how critical points for the linear optimization problem (7.1) vary under data perturbation. Furthermore, its existence forces the critical point $\bar{x} \in X$ for the vector $\bar{y} \in \mathbf{E}$ to be *nondegenerate*: $\bar{y}$ must lie not just in the normal cone $N_X(\bar{x})$, but in its relative interior — its interior relative to its span. It also forces a local form of prox-regularity [29], rather as in the Section 6, so for transparency we simply assume prox-regularity.

Theorem 7.1 (Identifiability, uniqueness, and sensitivity). *Suppose the set $X \subset \mathbf{E}$ is prox-regular at the point $\bar{x} \in X$. If X has an identifiable manifold $\mathcal{M}$ at $\bar{x}$ for the vector $\bar{y} \in \mathbf{E}$, then that manifold is locally unique. Indeed, for any sufficiently small neighborhood U of $\bar{y}$, the manifold $\mathcal{M}$ coincides locally around $\bar{x}$ with the set $N_X^{-1}(U)$. Furthermore, $\bar{x}$ must then be a nondegenerate critical point for $\bar{y}$.*

To illustrate, consider the case when the set X is a polyhedron. If $\bar{x}$ is a nondegenerate critical point for the problem $\sup_X \langle \bar{y}, \cdot \rangle$, then the set of maximizers (or in other words the face of X exposed by the vector $\bar{y}$) is an identifiable manifold at $\bar{x}$ for $\bar{y}$. We discuss more varied examples in the following sections.

A set X may easily have no identifiable manifold at the critical point $\bar{x}$ in question, even when X is closed, convex and semi-algebraic, and $\bar{x}$ is nondegenerate. An example is the set

$$X \;=\; \big\{(u, v, w) \in \mathbf{R}^3 : w^2 \geq u^2 + v^4,\; w \geq 0\big\},$$

at the point $\bar{x} = (0, 0, 0)$ for the vector $\bar{y} = (0, 0, -1)$. However, as we shall see shortly, at least for semi-algebraic examples such as this one, such behavior is unusual. Furthermore, as the next result [29] makes clear, the existence of an identifiable manifold has broad and powerful consequences for optimization.

Theorem 7.2 (Identifiability, active sets, and partial smoothness). *Suppose the set $X \subset \mathbf{E}$ is prox-regular at the point $\bar{x} \in X$, and has an identifiable manifold $\mathcal{M}$ there for the vector $\bar{y} \in \mathbf{E}$. Then the following properties hold:*

- **Smooth reduction:** *The graphs of the normal cone mappings N_X and $N_{\mathcal{M}}$ coincide around the point $(\bar{x}, \bar{y})$.*
- **Sharpness:** *The normal cone $N_X(\bar{x})$ spans the normal space $N_{\mathcal{M}}(\bar{x})$.*
- **Active set philosophy:** *For any small neighborhood V of $\bar{x}$, if the vector $y \in \mathbf{E}$ is near $\bar{y}$, then the two optimization problems of maximizing the linear function $\langle y, \cdot \rangle$ over the sets $X \cap V$ and $\mathcal{M} \cap V$ are equivalent.*
- **Second-order conditions:** *The rate of quadratic growth*

$$\liminf_{x \to \bar{x}} \frac{\langle \bar{y}, \bar{x} - x \rangle}{|\bar{x} - x|^2}$$

is independent of whether the limit is taken over $x \in X$ or $x \in \mathcal{M}$.

Given the multiple flavors of this result, some commentary is useful. Perhaps most striking is the "active set" result, which reduces the original optimization problem over the potentially nonsmooth and high-dimensional set X to the restricted optimization problem over the smooth and potentially lower-dimensional subset $\mathcal{M}$. Exactly this phenomenon drives the elimination of inequality constraints inherent in classical active set methods for optimization [61], and also the big reduction in dimension crucial to "sparse optimization" in contemporary machine learning and compressed sensing applications. In a huge recent literature, a particularly pertinent example is [80].

Underlying the active set assertion is the "smooth reduction" result that the mappings N_X and $N_{\mathcal{M}}$ graphically coincide, locally. Since $\mathcal{M}$ is a smooth manifold, its normal cone mapping is easy to understand through classical analysis. In particular, second-order properties like the negative definite condition in Theorem 6.3 (Strong regularity and quadratic growth), which may in general appear formidably abstract, now become purely classical [55]. For example, the $\liminf$ in the final second-order condition above, when computed over $\mathcal{M}$, simply involves a Hessian computation for the function $\langle \bar{y}, \cdot \rangle$ restricted to the manifold $\mathcal{M}$.

The "sharpness" property at the point $\bar{x}$ is geometric in essence: we call the set X *sharp* (or "V-shaped") there around the manifold $\mathcal{M}$. In [50], extending Wright's notion of an "identifiable surface" for active set methods in convex optimization [79], the set X is called *partly smooth* at the point $\bar{x}$ relative to the $\mathcal{C}^{(2)}$ manifold $\mathcal{M}$ when this sharpness property holds, the normal cone mapping N_X is continuous at $\bar{x}$ when restricted to $\mathcal{M}$, and *Clarke regularity* holds on $\mathcal{M}$. This latter property concerns *tangent* directions $z \in \mathbf{E}$ at any point $x \in X$ (limits of directions to nearby points in X): it requires $\langle y, z \rangle \leq 0$ for all normals $y \in N_X(x)$.

Partial smoothness is closely related to identifiability. In general, consider a point $\bar{x}$ in a set X and a proximal normal $\bar{y} \in N_X^p(\bar{x})$. On the one hand, suppose that the critical point $\bar{x}$ is nondegenerate for $\bar{y}$, and partial smoothness holds relative to a $\mathcal{C}^{(2)}$ manifold $\mathcal{M}$. In particular, X must then be Clarke regular at $\bar{x}$. However, if we strengthen this assumption slightly, from Clarke to prox-regularity, then $\mathcal{M}$ must be an identifiable manifold. On the other hand, suppose conversely that $\mathcal{M}$ is an identifiable manifold. As we have seen, $\bar{x}$ must then be nondegenerate, and furthermore a local version of partial smoothness must hold [29].

The existence of an identifiable manifold, as Theorem 7.2 makes clear, is a powerful property. Remarkably, according to the following result [32], for semi-algebraic optimization this property holds generically.

Theorem 7.3 (Generic identifiability). *Given any closed semi-algebraic set $X \subset \mathbf{E}$, there exists an integer K such that, for all vectors $y \in \mathbf{E}$ outside some negligible semi-algebraic subset of $\mathbf{E}$, the following properties hold. The linear optimization problem*

$$\sup_X \langle y, \cdot \rangle.$$

has no more than K local maximizers. At each local maximizer $x \in X$, the normal cone mapping is strongly metrically regular for y; there exists an identifiable manifold $\mathcal{M}$ at x for y, and the normal cone mappings N_X and $N_{\mathcal{M}}$ coincide around the point (x, y). Furthermore, x is a nondegenerate critical point, and X is sharp around $\mathcal{M}$ there: in other words, y lies in the interior of the normal cone $N_X(x)$ relative to its span, which is just the

normal space $N_{\mathcal{M}}(x)$. In addition, the following quadratic growth condition holds:

$$\liminf_{\substack{x' \to x \\ x' \in X}} \frac{\langle y, x - x' \rangle}{|x - x'|^2} > 0.$$

The key ingredients of the proof have appeared through our discussion. Generic strong metric regularity follows from Theorem 6.2, and in that case, the inverse image of a small neighborhood of y under the normal cone mapping N_X (or in other words the set of nearby approximately critical points) will generically comprise an identifiable manifold. The consequences then flow from Theorems 6.3, 7.1, and 7.2. For convex sets X, this result appeared in [5].

In this result, we can view the existence of an identifiable manifold in conjunction with nondegeneracy and the quadratic growth condition as comprising the natural "second-order sufficient conditions" for our optimization problem. Classically, the generic validity of such conditions has a long history, dating back to [70]. Here we have taken a fresh, abstract approach, assuming nothing about the structure of the problem beyond its concrete (semi-algebraic) nature.

8. Optimization over stable polynomials

We have argued that ideas of identifiable manifolds and active set methods in optimization merge seamlessly. Less standard, but an elegant computational illustration of the appearance of an identifiable manifold, is a problem of Blondel [4]. The original question (with generous prizes of Belgian chocolate) highlighted the difficulty of simultaneous plant stabilization in continuous-time control.

The crucial idea of stability in dynamical systems and control theory involves *stable* and *strictly stable* polynomials $p(z)$ (for the complex variable $z \in \mathbf{C}$): polynomials with all zeroes in the closed or open left half-planes respectively. Blondel's problem seeks stable polynomials p, q, r with real coefficients and satisfying

$$r(z) = (z^2 - 2\delta z + 1)p(z) + (z^2 - 1)q(z),$$

for a real parameter $\delta \in [0.9, 1)$. If $\delta = 1$, then $r(1) = 0$, so no solution exists.

An optimization approach to this problem in [10], for any fixed parameter value δ, varies a cubic polynomial p and scalar q to minimize numerically a real variable α under the condition that the two polynomials $z \mapsto p(z + \alpha)$ and $z \mapsto r(z + \alpha)$ are both stable. The numerical results in [10] strongly suggest that when $\delta \in [0.9, 0.96]$, the minimum value $\bar{\alpha}$ is negative, as required for stability. If, furthermore, δ is close to, and no larger than, the value $\bar{\delta} = \frac{1}{2}\sqrt{2 + \sqrt{2}} \approx 0.924$, then the optimal polynomials $\bar{p}$ and $\bar{r}$ are not only stable but have a persistent structure: $\bar{p}$ is strictly stable, and $\bar{r}$ is a multiple of the polynomial $z \mapsto (z - \bar{\alpha})^5$. This structure defines a manifold $\mathcal{M}$ in the space of variables (α, p, q, r), which, once divined numerically, leads to a solution to Blondel's problem in closed form for such δ. Not surprisingly, $\mathcal{M}$ is the identifiable manifold for our optimization problem.

Underlying this striking appearance of an identifiable manifold is a remarkable property of stable polynomials. To understand this property, we first identify monic polynomials p of degree n with vectors $\tilde{p}$ in the space $\mathbf{C}^n$ (with the usual inner product), via the correspondence $p(z) = z^n + \sum_{j<n} \tilde{p}_j z^j$, and thereby consider them as constituting a Euclidean

space. Within that space, we then consider the set of stable polynomials Δ_n. The basic variational geometry of this nonconvex set is challenging. Around any polynomial with a multiple imaginary zero, Δ_n is nonsmooth, and indeed, with a suitable interpretation, nonlipschitz. Notice, for example, that monic polynomials $p(z)$ near the polynomial z^n have zeroes whose dependence on the coefficient vector $\tilde{p}$ is nonlipschitz.

On the other hand, despite these structural challenges, the set of monic stable polynomials is certainly semi-algebraic. Theorem 7.3 (Generic identifiability) therefore implies the *generic* existence of an identifiable manifold around solutions of linear optimization problems over stable polynomials. However, the following beautiful result of Burke and Overton [13] holds not just generically, but *always*.

Theorem 8.1. *The set of monic stable polynomials of degree n is Clarke regular everywhere.*

The techniques of [13] (which treat regions more general than the left half-plane) show more. For any monic stable polynomial p, the normal cone $N_{\Delta_n}(p)$ depends on the "pattern" of imaginary zeroes of p (which we specify simply by listing the multiplicities of those zeroes as we move down the imaginary axis). Using the language of partial smoothness from Section 7, we arrive at the following result.

Theorem 8.2 (Partial smoothness of the stable polynomials). *Around any polynomial in the set of monic stable polynomials Δ_n, the subset of polynomials with the same pattern of imaginary zeroes constitute a manifold, with respect to which Δ_n is partly smooth.*

It is exactly this property that underlies the identifiable manifold in [10] for Blondel's problem. The set of stable n-by-n matrices — those whose eigenvalues all lie in the left half plane — enjoys parallel properties around any stable *nonderogatory* matrix (one whose eigenvalues all have geometric multiplicity one) [11, 12]. One explanation [51] is to note that the characteristic polynomial map from the space of matrices to monic polynomials has surjective derivative at any nonderogatory matrix, enabling a standard calculus rule.

9. An eigenvalue optimization example

We can be confident of the generic existence of an identifiable manifold for a semi-algebraic optimization problem, by Theorem 7.3 (Generic identifiability), under no assumptions whatsoever about the problem's presentation. Optimization algorithms sometimes reveal clues about the identifiable manifold as they proceed. For the polynomial stabilization example in Section 8, numerical results from a simple general-purpose nonsmooth optimization method point to the identifiable manifold, helped along by our understanding of the potential structure for such manifolds. The BFGS method that we discussed in Section 5 naturally accumulates identifiable manifold information as it nears an optimal solution.

Suppose the BFGS method for minimizing a function $f\colon \mathbf{R}^n \to \mathbf{R}$ converges to a local minimizer $\bar{x} \in \mathbf{R}^n$ at which there exists an identifiable manifold $\mathcal{M}$. By this, we mean that the set $\big\{\big(x, f(x)\big) : x \in \mathcal{M}\big\}$ is an identifiable manifold of the epigraph $\operatorname{epi} f$, at the point $\big(\bar{x}, f(\bar{x})\big)$, for the vector $(0, -1)$. We deduce a dichotomy: on the one hand, the restriction of f to the manifold $\mathcal{M}$ is smooth, and on the other hand, the sharpness condition in Theorem 7.2 (Identifiability, active sets, and partial smoothness) shows that the gradient ∇f jumps as we move orthogonally across $\mathcal{M}$. The inverse Hessian approximation H_k should reflect this dichotomy: a basis of eigenvectors spanning an approximation to the tangent space to

$\mathcal{M}$ at $\bar{x}$ corresponds to a well-scaled set of eigenvalues, whereas the eigenvectors spanning the orthogonal complement correspond to eigenvalues converging to zero. In numerical experiments on semi-algebraic Lipschitz functions, we indeed see exactly this behavior [53].

The following example from [1] is illuminating:

$$\inf\Big\{\prod_{i=1}^{q}\lambda_i(A\circ X) : X\in\mathbf{S}^p_+,\ X_{ii}=1 \text{ for all } i\Big\}.$$

Here, $\mathbf{S}^p_+$ denotes the cone of p-by-p positive semidefinite matrices, $A\in\mathbf{S}^p$ is a given data matrix, $\circ$ denotes the componentwise (Hadamard) matrix product, and λ_i denotes the ith largest eigenvalue (counted by multiplicity). It is not hard to frame this optimization problem as the unconstrained minimization of a suitable function f, expressed in terms of the nonsmooth nonconvex function $\prod_{i=1}^{q}\lambda_i$ on the space $\mathbf{S}^p$: see [53] for the modeling details.

The results from multiple runs of the BFGS method on an example with $p=20$ and $q=10$ are typical for eigenvalue optimization [53]. Generic symmetric matrices have no multiple eigenvalues, but optimal solutions of semidefinite programs (see [72]) and more general eigenvalue optimization problems usually do, precisely due to their identifiable manifolds. That is the case here: as observed in Section 5, the BFGS trial function values consistently converge linearly, and at termination, the nine eigenvalues $\lambda_6,\lambda_7,\lambda_8,\ldots,\lambda_{14}$ of the matrix $A\circ X$ are coalescing.

Given a permutation-invariant function $h\colon\mathbf{R}^p\to\mathbf{R}$, that function of the vector $\lambda(Z)\in\mathbf{R}^p$ with components the eigenvalues of a matrix variable $Z\in\mathbf{S}^p$ inherits many properties from h. One important example is convexity [48], a generalization of von Neumann's characterization of unitarily invariant matrix norms [75], but the list of such properties is extensive [49]. In particular [21], given a matrix $\bar{Z}\in\mathbf{S}^p$, if h has an identifiable manifold $\mathcal{M}$ at the point $\lambda(\bar{Z})$, then at $\bar{Z}$ the composite function $h\big(\lambda(\cdot)\big)$ has an identifiable manifold $\{Z\in\mathbf{S}^n : \lambda(Z)\in\mathcal{M}\}$.

The permutation-invariant function $h\colon\mathbf{R}^p\to\mathbf{R}$ here is $h(x)=\prod_{i=1}^{q}[x]_i$, where the map $x\mapsto[x]$ rearranges the components of the vector $x\in\mathbf{R}^p$ into nonincreasing order. For this function h we can easily check in the example that the set

$$\big\{x\in\mathbf{R}^{20} : [x]_5>[x]_6=[x]_7=\cdots=[x]_{14}>[x]_{15}\big\}$$

is an identifiable manifold. Hence

$$\big\{Z\in\mathbf{S}^{20} : \lambda_5(Z)>\lambda_6(Z)=\lambda_7(Z)=\cdots=\lambda_{14}(Z)>\lambda_{15}(Z)\big\}$$

is an identifiable manifold for our objective function $\prod_{i=1}^{10}\lambda_i$. Classical matrix analysis [46, p. 141] shows that this manifold of symmetric matrices with an eigenvalue of multiplicity nine has codimension $\frac{1}{2}9(9+1)-1=44$.

Examining the BFGS output [53] and counting the number of eigenvalues of the inverse Hessian approximations H_k that converge to zero reveals the answer 44 — exactly the codimension of the identifiable manifold at the optimal solution. To confirm, around the final iterate we can plot the behavior of the objective function along the eigenvectors of H_k. Sure enough, along those eigenvectors corresponding to the vanishing eigenvalues, the objective function is V-shaped; along other eigenvectors, it is smooth. To summarize, with no *a priori* input about the underlying structure of the problem, and no *a posteriori* interpretation, the BFGS method nonetheless accurately approximates the geometry of the identifiable manifold.

10. Identifiability and a prox-linear algorithm

We have argued that, independent of the presentation of an optimization problem, an identifiable manifold is typically there to be found, and is a powerful tool once known. However, the manner in which a particular algorithm profits from such knowledge will likely depend on the explicit structure of the underlying problem. The classical example is the active set methodology for optimization under inequality constraints, which considers equality-constrained subproblems based on an estimate of the "active set" of constraints — those that are tight at optimality. We end this survey with a discussion of a practical algorithm [54], designed for large-scale applications in areas such as machine learning, and well-suited to the application of identifiability.

Given two Euclidean spaces $\mathbf{E}$ and $\mathbf{F}$ and a closed set $Y \subset \mathbf{F}$, we consider optimization problems of the following form:

$$\inf_{x \in \mathbf{E}} \big\{ f(x) : g(x) \in Y \big\}, \tag{10.1}$$

where the functions $f \colon \mathbf{E} \to \mathbf{R}$ and $g \colon \mathbf{E} \to \mathbf{F}$ are $\mathcal{C}^{(2)}$ smooth. Crucially, we suppose that the set Y is, in some sense, simple. We define "simple" operationally: we assume that we can solve relatively easily *prox-linear subproblems* of the form

$$\inf_{d \in \mathbf{E}} \big\{ \tilde{f}(d) + \mu |d|^2 : \tilde{g}(d) \in Y \big\}, \tag{10.2}$$

for *affine* functions $\tilde{f} \colon \mathbf{E} \to \mathbf{R}$ and $\tilde{g} \colon \mathbf{E} \to \mathbf{F}$, and a *prox parameter* $\mu > 0$. In the algorithm we describe, $\tilde{f}$ and $\tilde{g}$ are the linear approximations to f and g at the current iterate x_k:

$$\tilde{f}(d) = f(x_k) + Df(x_k)d \ \text{ and } \ \tilde{g}(d) = g(x_k) + Dg(x_k)d.$$

Consider again the example of optimization under inequality constraints, when the set Y is just a positive orthant. The corresponding prox-linear subproblem reduces to projection onto a polyhedron, a relatively easy problem computationally. Simpler still is the l_1-constrained least squares problem

$$\inf_{x \in \mathbf{R}^n} \big\{ |Ax - b|^2 : |x|_1 \leq \tau \big\},$$

for given $\tau > 0$, used to find sparse approximate solutions to huge linear systems $Ax = b$ in popular procedures such as LASSO and LARS [15, 23, 34, 71]. Corresponding prox-linear subproblems at the point x have the form

$$\inf_{d \in \mathbf{R}^n} \big\{ 2\langle Ax - b, Ad \rangle + \mu |d|^2 : |x + d|_1 \leq \tau \big\}. \tag{10.3}$$

This problem reduces to projection onto the l_1-ball, for which very fast algorithms are available: simple $O(n \log n)$ methods appear in [33, 73], and [33] describes an approach in expected linear time. (The computational simplicity of the singly-constrained convex program (10.3) is not surprising: its Lagrangian is separable in the components d_i, so can be minimized in linear time.) The nuclear-norm-constrained least squares approach for low-rank matrix equations is similar [14]. This time the corresponding subproblem is projection onto the nuclear-norm-ball (consisting of matrices whose singular values sum to at most one), which again is relatively easy: we simply replace the vector of singular values appearing in the singular value decomposition by its projection onto the l_1-ball.

Returning to the general problem (10.1), we consider a local minimizer $\bar{x}$ at which the set Y is prox-regular and satisfies the following standard constraint qualification:

$$\text{span } N_Y\big(g(\bar{x})\big) \cap \text{Null}\big(Dg(\bar{x})^*\big) \;=\; \{0\}. \tag{10.4}$$

This condition implies that $\bar{x}$ must satisfy the natural *first-order optimality condition*: there exists a *Lagrange multiplier* $y \in \mathbf{F}$ (in fact unique) such that

$$y \in N_Y\big(g(\bar{x})\big) \;\text{ and }\; Df(\bar{x}) + Dg(\bar{x})^* y = 0. \tag{10.5}$$

Furthermore, for any point $x \in \mathbf{E}$ near $\bar{x}$, if the prox parameter μ is large enough, then the prox-linear subproblem (10.2) has a unique small local minimizer $d(x)$, and in fact $d(x) = O(|x - \bar{x}|)$.

The basic structure of the algorithm we describe is standard in optimization. The prox parameter μ controls the size of the trial step suggested by the prox-linear subproblem. When μ is large enough, we can correct the trial step to generate a reasonable fraction of the improvement predicted by linearization. If that proves impossible, we retrench, rejecting the trial step and increasing μ.

To be more precise, suppose the current iterate is $x \in \mathbf{E}$, and the current value of the prox parameter is $\mu > 0$. We first calculate the trial step $d = d(x)$, the appropriate local minimizer for the prox-linear subproblem (10.2), so in particular

$$g(x) + Dg(x)d \in Y \tag{10.6}$$

holds. We then calculate the new iterate $x^+ \in \mathbf{E}$ by trying to *correct* the trial point $x + d$, aiming at three conditions. First, the correction should be not too large relative to the step:

$$\big|x^+ - (x + d)\big| \leq \frac{1}{2}|d|.$$

Secondly, the new iterate should be feasible: $g(x^+) \in Y$. Thirdly, the actual decrease in the objective should be at least a reasonable fraction of that predicted by linearization:

$$\frac{f(x) - f(x^+)}{f(x) - f(x + d)} \geq \frac{1}{2}.$$

Assuming μ is sufficiently large, the constraint qualification (10.4) ensures that such a correction x^+ exists. If we find it, we accept it as our new current iterate and proceed; if not, we reject it, double μ, and try the whole process again. A standard argument shows a rudimentary convergence result: any limit point of the sequence of iterates must satisfy the first-order optimality condition.

The ideas behind this algorithm date back three decades [9, 37]. An implementable version in general must overcome two hurdles. The first — that the prox-linear subproblem may have several local minimizers — may arise, but only for nonconvex sets Y. The second concerns the correction mechanism, which we leave unspecified. When the map g is linear, or in particular just the identity, the algorithm is workable without correction. The algorithm we have described for the special case $\inf_Y f$, for closed convex Y and smooth f (which covers l_1-constrained least squares, for example), is closely related to the successful SPARSA code for compressed sensing [78]. Some kind of correction step is crucial when the map g is nonlinear, no matter how large the prox parameter μ. In particular, the linearized constraint

(10.6) does not guarantee the feasibility condition $g(x+d) \in Y$. Even when the trial step is feasible, we may want to enhance it using second-order information, leading us back to the idea of identifiability.

The basic prox-linear algorithm that we have described is versatile: it is often simple to implement and applicable to large-scale problems. In general, however, its convergence is slow. For example, consider unconstrained minimization of a strictly convex quadratic: in this simple case, $f(x) = \langle x, Ax\rangle$ for a positive-definite self-adjoint map $A\colon \mathbf{E} \to \mathbf{E}$, the map g is just the identity, and the set Y is just $\mathbf{E}$. The prox-linear algorithm then becomes the method of steepest descent for f with a fixed step size, an algorithm that, as we observed in the introduction, converges linearly but slowly when the map A is ill-conditioned. If our algorithm can readily access second-order information, we might hope to accelerate convergence.

So far we have supposed that the set Y is simple enough to render the prox-linear subproblems relatively easy. Now assume we know more, namely the structure of the set's identifiable manifolds. For example, the identifiable manifolds of the l_1-ball in $\mathbf{R}^n$ are simply its interior along with the sets of vectors x with norm $|x|_1 = 1$ and constant sign pattern $(\operatorname{sgn} x_i)$, where $\operatorname{sgn}\gamma = \gamma/|\gamma|$, or zero if $\gamma = 0$.

This structural information about the set Y allows us to impose a *second-order optimality condition* of the kind guaranteed generically by Theorem 7.3 (Generic identifiability). Specifically, consider any point $\bar{x} \in \mathbf{E}$ satisfying feasibility ($g(\bar{x}) \in Y$), the constraint qualification (10.4), and the first-order optimality condition (10.5), and now suppose furthermore that Y has an identifiable manifold $\mathcal{M}$ at the point $g(\bar{x})$ for the normal vector y and that the objective f grows quadratically on the manifold $g^{-1}(\mathcal{M})$ around $\bar{x}$. This latter condition is classical, amounting to the requirement that the Hessian of the Lagrangian function $f+\langle y, g\rangle$ at $\bar{x}$ be positive definite on the tangent space to $\mathcal{M}$ at $\bar{x}$.

From the second-order optimality condition we deduce powerful consequences. First, around the critical point $\bar{x}$, the objective f must in fact grow at least quadratically not just on the identifiable manifold $\mathcal{M}$ but on the whole set Y. Secondly, initiated nearby, the prox-linear algorithm must converge to $\bar{x}$. Thirdly, the sequence of trial iterates $g(x_k)+Dg(x_k)d_k$ in Y generated by the prox-linear subproblems is asymptotically critical for the Lagrange multiplier y, and hence eventually lies in $\mathcal{M}$. The algorithm thus *identifies* $\mathcal{M}$, in principle allowing an eventual reduction of the original optimization problem to the classical equality-constrained problem $\inf\{f(x) : g(x) \in \mathcal{M}\}$, and thereby opening up the possibility of second-order methods and accelerated convergence, as in [80] for the LASSO problem.

11. Afterthoughts and acknowledgements

Variational analysis and nonsmooth optimization deserve a wide audience. A flourishing toolkit of elegant theory for several decades, the discipline's more computational impact is only now coming into focus. In its full generality (skirted here) the field can seem at first formidably technical. However, as this essay has tried to emphasize, the core ideas — the normal cone and metric regularity, for example — are intuitive and powerful in both theory and algorithms. Semi-algebraic variational analysis makes for an illuminating concrete testing ground for the theory. The reach of variational analysis in applications ranges from its historical roots in optimal control and the calculus of variations, through more recent domains such as eigenvalue optimization and robust control, and on to burgeoning areas like

compressed sensing and machine learning. The field is thriving.

The material in this essay strongly reflects what I have tried to learn from the many co-authors and mentors with whom I have been lucky enough to work. Among them, I would especially like to mention Jon Borwein (who taught me variational analysis), Jim Burke and Michael Overton (my enthusiastic companions watching theory made manifest on a computer screen), Jim Renegar (an inspiring source of encouragement), Jérôme Bolte and Aris Daniilidis (with whom I first explored the semi-algebraic world), and most recently Dima Drusvyatskiy and Alex Ioffe. Thanks too to Asen Dontchev, Mike Todd, and Steve Wright for their broad support, and their helpful suggestions on this manuscript.

The author is grateful to the Dipartimento di Ingegneria Informatica Automatica e Gestionale at the Università di Roma La Sapienza for its hospitality during the writing of this paper.

References

[1] K. Anstreicher and J. Lee, *A masked spectral bound for maximum-entropy sampling*, In A. Bucchianico, A. Läuter, and H. P. Wynn, editors, MODA 7 – Advances in Model-Oriented Design and Analysis, Springer, Berlin, 2004, pp. 1–10.

[2] S. Banach, *Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales*, Fundamenta Mathematicae **3** (1922), 133–181.

[3] H.H. Bauschke and J.M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Review **38** (1996), 367–426.

[4] V. Blondel, *Simultaneous Stabilization of Linear Systems*, Springer, Berlin, 1994.

[5] J. Bôlte, A. Daniilidis, and A. S. Lewis, *Generic optimality conditions for semi-algebraic convex programs*, Mathematics of Operations Research **36** (2011), 55–70.

[6] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.

[7] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer, New York, second edition, 2006.

[8] J. M. Borwein and Q. J. Zhu, *Techniques of Variational Analysis*, Springer, New York, 2005.

[9] J. V. Burke, *Descent methods for composite nondifferentiable optimization problems*, Mathematical Programming **33** (1985), 260–279.

[10] J. V. Burke, D. Henrion, A. S. Lewis, and M. L. Overton, *Stabilization via nonsmooth, nonconvex optimization*, IEEE Transactions on Automatic Control **51** (2006), 1760–1769.

[11] J. V. Burke, A. S. Lewis, and M. L. Overton, *Optimal stability and eigenvalue multiplicity*, Foundations of Computational Mathematics **1** (2001), 205–225.

[12] J. V. Burke and M. L. Overton, *Variational analysis of non-Lipschitz spectral functions*, Mathematical Programming **90** (2001), 317–352.

[13] ———, *Variational analysis of the abscissa mapping for polynomials*, SIAM Journal on Control and Optimization **39** (2001), 1651–1676.

[14] J.-F. Cai, E. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization **20** (2010), 1956–1982.

[15] E. J. Candès and T. Tao, *Near-optimal signal recovery from random projections: universal encoding strategies*, IEEE Transactions on Information Theory **52** (2007), 5406–5425.

[16] F.H. Clarke, *Necessary Conditions for Nonsmooth Problems in Optimal Control and the Calculus of Variations*, PhD thesis, University of Washington, Seattle, 1973.

[17] ———, *Generalized gradients and applications*, Transactions of the American Mathematical Society **205** (1975), 247–262.

[18] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[19] M. Coste, *An Introduction to O-minimal Geometry*, RAAG Notes, 81 pages, Institut de Recherche Mathématiques de Rennes, 1999.

[20] ———, *An Introduction to Semialgebraic Geometry*, RAAG Notes, 78 pages, Institut de Recherche Mathématiques de Rennes, 2000.

[21] A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis, *Orthogonal invariance and identifiability*, SIAM Journal on Optimization, 2014. To appear.

[22] J.W. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numerische Mathematik **51** (1987), 251–289.

[23] D. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory **52** (2006), 1289–1306.

[24] A. L. Dontchev and H. Frankowska, *On derivative criteria for metric regularity*, In Computational and Analytical Mathematics, Springer Proceedings in Mathematics and Statistics, Springer, New York, 2013, pp. 365–374.

[25] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar, *The radius of metric regularity*, Transactions of the American Mathematical Society **355** (2003), 493–517.

[26] A. L Dontchev and R. T. Rockafellar, *Implicit Functions and Solution Mappings: a View from Variational Analysis*, Springer, New York, 2009.

[27] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, *The dimension of semi-algebraic subdifferential graphs*, Nonlinear Analysis **75** (2012), 1231–1245.

[28] ———, *Alternating projections and coupling slope*, 2014. Preprint. arXiv:1401.7569.

[29] D. Drusvyatskiy and A. S. Lewis, *Optimality, identifiability, and sensitivity*, Mathematical Programming, 2013. DOI 10.1007/s10107-013-0730-4.

[30] _____, *Semi-algebraic functions have small subdifferentials*, Mathematical Programming, Series B **140** (2013), 5–29.

[31] _____, *Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential*, SIAM Journal on Optimization **23** (2013), 256–267.

[32] _____, *Strong regularity of semi-algebraic mappings*, 2014. Preprint.

[33] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient projections onto the l_1-ball for learning in high dimensions*, In Proceedings of the 25th International Conference on Machine Learning, Helsinki, 2008.

[34] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Annals of Statistics **32**(2) (2004), 407–499.

[35] I. Ekeland, *On the variational principle*, Journal of Mathematical Analysis and Applications **47** (1974), 324–353.

[36] R. Fletcher, *A new variational result for quasi-Newton formulae*, SIAM Journal on Optimization **1** (1991), 18–21.

[37] R. Fletcher and E. Sainz de la Maza, *Nonlinear programming and nonsmooth optimization by successive linear programming*, Mathematical Programming **43** (1989), 235–256.

[38] K. M. Grigoriadis and R. E. Skelton, *Low-order control design for LMI problems using alternating projection methods*, Automatica **32** (1996), 1117–1125.

[39] A. Grothendieck, *Esquisse d'un programme*, In L. Schneps and P. Lochak, editors, Geometric Galois Actions, volume 1. Cambridge University Press, Cambridge, U.K., 1997. London Mathematical Society Lecture Note Series 242.

[40] R. A. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, U.K., 1985.

[41] A. D. Ioffe, *Metric regularity and subdifferential calculus*, Russian Mathematical Surveys **55** (2000), 501–558.

[42] _____, *A Sard theorem for tame set-valued mappings*, Journal of Mathematical Analysis and Applications **335** (2007), 882–901.

[43] _____, *On the theory of subdifferentials*, Advances in Nonlinear Analysis **1** (2012), 47–120.

[44] N. H. Josephy, *Newton's method for generalized equations and the PIES energy model*, PhD thesis, Dept. of Industrial Engineering, University of Wisconsin-Madison, 1979.

[45] A. Y. Kruger and B. S. Mordukhovich, *Extremal points and the Euler equation in nonsmooth analysis*, Doklady Akademia Nauk BSSR (Belorussian Academy of Sciences) **24** (1980), 684–687.

[46] P. D. Lax, *Linear Algebra*, Wiley, New York, 1997.

[47] J. M. Lee, *Introduction to Smooth Manifolds*, Springer, New York, 2003.

[48] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM Journal on Optimization **6** (1996), 164–177.

[49] ———, *Nonsmooth analysis of eigenvalues*, Mathematical Programming **84** (1999), 1–24.

[50] ———, *Active sets, nonsmoothness and sensitivity*, SIAM Journal on Optimization **13** (2003), 702–725.

[51] ———, *Eigenvalues and nonsmooth optimization*, In L.M. Pardo, A. Pinkus, E. Suli, and M.J. Todd, editors, Foundations of Computational Mathematics, Santander 2005, Cambridge University Press, Cambridge, U.K., 2005, pp. 208–229.

[52] A. S. Lewis, D. R. Luke, and J. Malick, *Local linear convergence for alternating and averaged nonconvex projections*, Foundations of Computational Mathematics **3** (2009), 485–513.

[53] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Mathematical Programming **141** (2013), 135–163.

[54] A. S. Lewis and S. J. Wright, *A proximal method for composite minimization*, 2008. Preprint. arXiv:0812.0423v1.

[55] A. S. Lewis and S. Zhang, *Partial smoothness, tilt stability, and generalized Hessians*, SIAM Journal on Optimization **23** (2013), 74–94.

[56] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM Journal on Control and Optimization **30** (1992), 408–425.

[57] B. S. Mordukhovich, *Maximum principle in the problem of time optimal response with nonsmooth constraints*, Journal of Applied Mathematics and Mechanics **40** (1976), 960–969.

[58] ———, *Variational Analysis and Generalized Differentiation, I: Basic Theory; II: Applications*, Springer, New York, 2006.

[59] B. S. Mordukhovich and A. Y. Kruger, *Necessary optimality conditions in the problem of terminal control with nonfunctional constraints. (Russian)*, Doklady Akademia Nauk BSSR (Belorussian Academy of Sciences) **20** (1976), 1064–1067.

[60] Y. E. Nesterov and A. S. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[61] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, second edition, 2006.

[62] T. Pennanen, *Local convergence of the proximal point algorithm and multiplier methods without monotonicity*, Mathematics of Operations Research **27** (2002), 170–191.

[63] R. A. Poliquin and R. T. Rockafellar, *Tilt stability of a local minimum*, SIAM Journal on Optimization **8** (1998), 287–299.

[64] R. A. Poliquin, R. T. Rockafellar, and L. Thibault, *Local differentiability of distance functions*, Transactions of the American Mathematical Society **352** (2000), 5231–5249.

[65] J. Renegar, *Condition numbers, the barrier method, and the conjugate gradient method*, SIAM Journal on Optimization **6** (1996), 879–912.

[66] S. M. Robinson, *Strongly regular generalized equations*, Mathematics of Operations Research **5** (1980), 43–62.

[67] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[68] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[69] A. Sard, *The measure of the critical values of differentiable maps*, Bulletin of the American Mathematical Society **48** (1942), 883–890.

[70] J. E. Spingarn and R. T. Rockafellar, *The generic nature of optimality conditions in nonlinear programming*, Mathematics of Operations Research **4** (1979), 425–430.

[71] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society B **58** (1996), 267–288.

[72] M. J. Todd, *Semidefinite optimization*, Acta Numerica **10** (2001), 515–560.

[73] E. van den Berg and M. P. Friedlander, *Probing the Pareto frontier for basis pursuit solutions*, SIAM Journal on Scientific Computing **31** (2008), 890–912.

[74] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Mathematics Journal **84** (1996), 497–540.

[75] J. von Neumann, *Some matrix inequalities and metrization of matric-space*, Tomsk University Review **1** (1937), 286–300. In: Collected Works, Pergamon, Oxford, 1962, Volume IV, 205-218.

[76] ———, *Functional Operators*, volume II. Princeton University Press, Princeton, NJ, 1950. Reprint of notes distributed in 1933.

[77] H. Whitney, *A function not constant on a connected set of critical points*, Duke Mathematics Journal **1** (1935), 514–517.

[78] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Transactions on Signal Processing **57** (2009), 2479–2493.

[79] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM Journal on Control and Optimization **31** (1993), 1063–1079.

[80] ———, *Accelerated block-coordinate relaxation for regularized optimization*, SIAM Journal on Optimization **22** (2012), 159–186.

School of Operations Research and Information Engineering, Cornell University, Ithaca NY 14853, USA

E-mail: adrian.lewis@cornell.edu

Carleman estimates, results on control and stabilization for partial differential equations

Luc Robbiano

Abstract. In this survey we give some results based on Carleman estimates. We recall the classical uniqueness result based on interior Carleman estimate. We give Carleman estimate up the boundary useful for the applications. The main applications are, approximative control for wave equation, null control for heat equation, stabilization for wave equation for an interior damping or for a boundary damping and local energy decay for wave equation in exterior domain.

Mathematics Subject Classification (2010). 35A02, 35Q93, 93B05, 93D15.

Keywords. Carleman estimates, control, null control, stabilization, exterior problem.

1. Introduction

The Carleman estimates play an important role in the study of the uniqueness problem for operator with non analytic coefficients since Carleman [37]. He proved estimates in L^1 norm but the most part of Carleman estimates was later proven in L^2 norm. Nevertheless Carleman estimates in L^p norm was proven to study unique continuation (see [16, 43, 50, 79, 129, 132, 140]) or strong uniqueness (see [4, 5, 12, 38, 39, 74, 80, 83, 117, 118, 130, 141]) for operators with coefficients in L^q, where p is related to q, but it is not the main subject of this survey.

The general results on unique continuation was obtained first by Calderón [34] and Hörmander [64] who has found two conditions almost necessary to obtain Carleman estimates. The first, the principal normality (this condition was precise later by Lerner [94]) concerns the operators with complex valued principal symbols, the second is the pseudo-convexity condition. We may find the proofs of Calderón result and the Hörmander's result in the book [66, Chapter 28]. We shall recall these results in section 2 and some references on non uniqueness results.

For the applications to control problem, stabilisation and other related problems, we need Carleman estimates up the boundary. In section 3 we shall give two kinds of Carleman estimates up the boundary, first when the norms of boundary data and the operator applied on a function estimate the interior norm of the function, second when the boundary data are estimated only by the operator applied on the function. In second case we need boundary conditions, for instance Dirichlet or Neumann boundary conditions.

In section 4 we shall give some applications using Carleman estimates. Using a Fourier Bros Iagolnitzer (FBI) transform in time variable, we can transform a wave equation in

Proceedings of the International Congress of Mathematicians, Seoul, 2014

elliptic equation. With Carleman estimate for elliptic operators, this allows to prove that the wave equation is always approximatively controllable and we can give a quantitative estimate of this result. Actually we can estimate the cost function, i.e. the L^2 norm of the control, to reach a small ball around the target.

For heat equation we shall give results on null controllability, first by a spectral approach and second by the method developed by Fursikov and Imanuvilov using a new kind of Carleman estimate adapted to the heat equation.

We give results on stabilization for wave equation, first if the damping is localized in the domain and second if the damping acts on the boundary. In both cases we estimate the decay of the energy. In the same spirit we give some results on the decay of local energy for wave equation in exterior domain.

Carleman estimates was applied to others problems. At the end of this section we shall give some references on such applications for degenerate parabolic equations, equations obtain by discretization of partial differential equation, stochastic equations, optimal control in time and inverse problems.

2. Local continuation results and interior Carleman estimates

The traditional classification in partial differential equation, hyperbolic equations, elliptic equations, parabolic equations, etc. is not relevant for local unique continuation. The two important notions are the principal normality and the pseudo-convexity. We shall give the definitions of these notions, next we shall give the results on local unique continuation.

We recall some notations. Let $P = \sum_{|\alpha|\leq m} a_\alpha(x)D^\alpha$ be a differential operator where $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n$, $m \in \mathbb{N}^*$, $x \in \mathbb{R}^n$, $D^\alpha = D_1^{\alpha_1} \cdots D_n^{\alpha_n}$ and $D_j = -i\partial_{x_j}$. The functions a_α are complex valued, defined and $\mathscr{C}^\infty$ in a neighborhood W of $x_0 \in \mathbb{R}^n$. For $\xi = (\xi_1, \cdots, \xi_n) \in \mathbb{R}^n$, we denote by $p(x,\xi) = \sum_{|\alpha|=m} a_\alpha(x)\xi^\alpha$, where $\xi^\alpha = \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n}$, the principal symbol of P. Let φ be a real valued function, $\mathscr{C}^\infty$ in W, we assume $d\varphi(x) \neq 0$ for $x \in W$. We define the surface $S = \{x \in W,\ \varphi(x) = \varphi(x_0)\}$.

Definition 2.1 (Principal normality). We say that P or p are principally normal if there exists $C > 0$ such that

$$|\{p, \bar{p}\}(x,\xi)| \leq C|p(x,\xi)||\xi|^{m-1}, \tag{2.1}$$

for all $x \in W$ and $\xi \in \mathbb{R}^n \setminus 0$.

Remark 2.2. If we denote $p = p_1 + ip_2$, where p_j are real valued, we have $\{p, \bar{p}\}(x,\xi) = 2i\{p_2, p_1\}(x,\xi)$ then we can replace (2.1) by the condition $|\{p_1, p_2\}(x,\xi)| \leq C(|p_1(x,\xi)| + |p_2(x,\xi)|)|\xi|^{m-1}$.

Remark 2.3. For operators with real coefficients in principal symbol, this definition is empty since $\{p, \bar{p}\}(x,\xi) = 0$

Definition 2.4 (pseudo-convexity). We say that S is strongly pseudo-convex with respect to P at x_0 if

$$\forall \xi \in \mathbb{R}^n \setminus \{0\},\ p(x_0,\xi) = 0, \{p,\varphi\}(x_0,\xi) = 0 \Rightarrow \mathrm{Re}\{\bar{p}, \{p,\varphi\}\}(x_0,\xi) > 0 \tag{2.2}$$

and

$$\forall \xi \in \mathbb{R}^n \setminus \{0\},\ \forall \tau > 0,\ p(x_0, \xi + i\tau d\varphi(x_0)) = 0, \{p,\varphi\}(x_0, \xi + i\tau d\varphi(x_0)) = 0$$

$$\Rightarrow \mathrm{Im}\{\bar{p}(x,\xi - i\tau d\varphi(x)), p(x,\xi + i\tau d\varphi(x))\} > 0 \text{ at } (x,\xi) = (x_0,\xi). \tag{2.3}$$

Remark 2.5. If we denote by $p_\varphi(x,\xi) = p(x,\xi + i\tau d\varphi(x))$, where τ is a parameter, we have

$$\{\bar{p}_\varphi, p_\varphi\}(x,\xi) = \{\bar{p}(x,\xi - i\tau d\varphi(x)), p(x,\xi + i\tau d\varphi(x))\}.$$

In particular this proves that the condition (2.3) is invariant by change of variables.

Remark 2.6. We can replace φ by other function defining the same oriented surface S, that is $\phi(x) = g(x)\varphi(x)$ with $g(x) > 0$, then the conditions (2.2) and (2.3) are invariant under this change. This mean that the strongly pseudo-convexity is a geometrical condition on S an oriented sub-manifold of co-dimension one.

Remark 2.7. The conditions (2.2) and (2.3) are open then the conditions are true in a neighborhood of x_0.

Remark 2.8. If p is principally normal then the limit of condition (2.3) when τ goes to 0 gives the condition (2.2). Actually the definition of strong pseudo-convexity make sense only for principal normal symbol.

Remark 2.9. Let $Q(\tau) = p(x_0, \xi + \tau d\varphi(x))$, for $\xi \neq 0$. If the roots of Q are simple i.e. if $Q(\tau) = 0$ implies $Q'(\tau) \neq 0$, this is equivalent to $p(x_0, \xi + i\tau d\varphi(x_0)) = 0$ and $\{p,\varphi\}(x_0, \xi + i\tau d\varphi(x_0)) \neq 0$. Then if the root of Q are simple S is strongly pseudo-convex with respect P. In this case the condition is also satisfied for $-\varphi$ this means that we can change the orientation of S. The uniqueness result given by Hörmander (see Theorem 2.13 below) contains the Calderón's result [34] but does not contain all known results where the roots of Q are assumed smooth and eventually multiple. See Hörmander [66, Chapter 28, section 1] for uniqueness results under smoothness assumptions on roots and Zuily [143, Chapter 2] .

Remark 2.10. There exists a lot of results on non-uniqueness. The results have the following form there exist u and a such that $Pu + au = 0$, $u = 0$ on $\varphi(x) < \varphi(x_0)$ in a neighborhood of x_0 and x_0 is in the support of u. We can find such results in [2, 3, 6, 119, 124, 125]. Essentially in all cases where principal normality or strong pseudo-convexity are not satisfied in a strong sense there exists a non uniqueness result in the sense given above.

Remark 2.11. Very few results exist under weak assumption of pseudo-convexity, see results in this direction in [13, 15, 96, 126] and [66, Th. 28.4.3]).

Remark 2.12. If p is of second order with real coefficients, p is principally normal and does not have double complex roots. Then the pseudo-convexity condition is only given by (2.2). In this case, the pseudo-convexity condition can be described in term of bicharacteristic. Let $(x(s),\xi(s))$ be the solution of

$$\begin{cases} \dot{x}(s) = \partial_\xi p(x(s),\xi(s)) \\ \dot{\xi}(s) = -\partial_x p(x(s),\xi(s)) \\ x(0) = x_0, \quad \xi(0) = \xi_0, \end{cases}$$

that is $(x(s),\xi(s))$ is the integral curve of the vector field $H_p = \partial_\xi p(x,\xi)\partial_x - \partial_x p(x,\xi)\partial_\xi$. Let $g(s) = \varphi(x(s))$, we have

$$\dot{g}(s) = \partial_x\varphi(x(s))\dot{x}(s) = \partial_x\varphi(x(s))\partial_\xi p(x(s),\xi(s)) = \{p,\varphi\}(x(s),\xi(s)).$$

The second derivative of g gives

$$\begin{aligned}\ddot{g}(s) &= \partial_x\{p,\varphi\}(x(s),\xi(s))\dot{x}(s) + \partial_\xi\{p,\varphi\}(x(s),\xi(s))\dot{\xi}(s)\\ &= \partial_x\{p,\varphi\}(x(s),\xi(s))\partial_\xi p(x(s),\xi(s)) - \partial_\xi\{p,\varphi\}(x(s),\xi(s))\partial_x p(x(s),\xi(s))\\ &= \{p,\{p,\varphi\}\}(x(s),\xi(s)).\end{aligned}$$

If we have $p(x_0,\xi_0) = \{p,\varphi\}(x_0,\xi_0) = 0$, then

$$\varphi(x(s)) = \varphi(x_0) + (s^2/2)\{p,\{p,\varphi\}\}(x_0,\xi_0) + O(s^3)$$

The pseudo-convexity condition is equivalent to say, the tangent bicharacteristics to S have a contact to order 2 with S and stay in $\{x,\ \varphi(x) \geq \varphi(x_0)\}$.

Now we can give the statement of the unique continuation theorem.

Theorem 2.13. *Let W be an open set and $x_0 \in W$. We assume P principally normal and S strongly pseudo-convex with respect P at x_0 then there exists V an open subset of W with $x_0 \in V$ such that for all $u \in \mathscr{C}^\infty(W)$*

$$Pu = 0 \text{ in } W \text{ and } u = 0 \text{ in } \{x \in W, \varphi(x) > \varphi(x_0)\} \Rightarrow u = 0 \text{ in } V \tag{2.4}$$

Remark 2.14. The assumption $u \in \mathscr{C}^\infty(W)$ is not a restriction. If the coefficients of P are smooth. Under the assumption given in Theorem 2.13, a distribution u satisfying the assumption is necessary $\mathscr{C}^\infty$ in a neighborhood of x_0 (see [64, Theorem 8.8.1]).

The proof of Theorem 2.13 is given by a Carleman estimate stated now.

Theorem 2.15. *Let P be principally normal, and $\phi \in \mathscr{C}^\infty(W)$ such that $d\phi(x_0) \neq 0$,*

$$\forall \xi \in \mathbb{R}^n \setminus \{0\},\ p(x_0,\xi) = 0 \Rightarrow \mathrm{Re}\{\bar{p}\{p,\phi\}\}(x_0,\xi) > 0 \tag{2.5}$$

and

$$\begin{aligned}&\forall \xi \in \mathbb{R}^n \setminus \{0\},\ \forall \tau > 0,\ p(x_0,\xi + i\tau d\phi(x_0)) = 0,\\ &\Rightarrow \mathrm{Im}\{\bar{p}(x,\xi - i\tau d\phi(x)), p(x,\xi + i\tau d\phi(x))\} > 0 \text{ at } (x,\xi) = (x_0,\xi).\end{aligned} \tag{2.6}$$

Then there exist V a neighborhood of x_0, $C > 0$ and $\tau_0 > 0$ such that for all $u \in \mathscr{C}_0^\infty(V)$ and all $\tau \geq \tau_0$

$$\sum_{|\alpha| \leq m-1} \tau^{2m-2|\alpha|-1} \|e^{\tau\phi(x)} D^\alpha u(x)\|^2 \leq C \|e^{\tau\phi(x)} Pu(x)\|^2. \tag{2.7}$$

Remark 2.16. The assumption to obtain uniqueness are invariant by change of variables and by change of defining function of the oriented surface. The estimate (2.7) is invariant by change of variables but not by multiplication of function ϕ by a positive function. We can remark that the assumptions (2.5) and (2.6) are not invariant by change of function ϕ. Actually is φ satisfies (2.2) and (2.3) then $\phi = e^{\lambda\varphi}$ satisfies (2.5) and (2.6) for λ large enough. To prove Theorem 2.13 we need to convexify the surface S. More precisely if φ satisfies (2.2) and (2.3) then $\varphi(x) - \varepsilon|x - x_0|^2$ satisfies also (2.2) and (2.3) maybe on a smaller neighborhood of x_0 because these conditions are open. Then $\phi(x) = e^{\lambda(\varphi(x) - \varepsilon|x-x_0|^2)}$ satisfies (2.5) and (2.6) for λ large enough.

Remark 2.17. In the estimate (2.7) we can add to P lower order term with bounded complex valued coefficients as these terms can be estimated by the left hand side. In particular the uniqueness results are also true for these operators. If the principal symbol has non smooth coefficients, there are positive results for instance in Hörmander [64] but also counter-examples see [116], [102] and [65].

Remark 2.18. The condition (2.6) can be interpreted as a sub-elliptic condition on the principal symbol $p(x, \xi + i\tau d\phi(x))$, and (2.7) as a sub-elliptic estimate with lost of one half derivative. In Lerner [95] we can find Carleman estimates under sub-elliptic conditions of superior order with related lost of derivative.

Remark 2.19. We can find some results in literature on Carleman estimate for systems. When the determinant of principal symbol satisfied assumptions (2.5) and (2.6), we can deduce Carleman estimate from previous result (see [64, Chapter 8, and (3.8.5)]). For elasticity system see [42] and [10], and for Stokes system see [51].

3. Boundary Carleman estimates

For the applications we need Carleman estimates up the boundary. Here we only consider operator P of order 2, elliptic. Results in more general cases can be found in Tataru [136].

For elliptic operator of order 2 with real coefficients, the principal normality condition and the condition (2.5) are trivially satisfied. The sub-elliptic condition (2.6) takes the following form

$$\begin{aligned}&\forall \xi \in \mathbb{R}^n \setminus \{0\},\ \forall \tau > 0,\ p(x_0, \xi + i\tau d\phi(x_0)) = 0,\\ &\Rightarrow \mathrm{Im}\{p(x, \xi - i\tau d\phi(x), p(x, \xi + i\tau d\phi(x)\} > 0 \text{ at } (x, \xi) = (x_0, \xi).\end{aligned} \tag{3.1}$$

In this section we assume P defined in a neighborhood of Ω, a bounded connected open set in $\mathbb{R}^n$ with $\mathscr{C}^\infty$ boundary. First we give local result in a neighborhood of a point $x_0 \in \partial\Omega$. The Carleman estimate up the boundary shall be given for smooth functions up the boundary compactly supported. To be precise, let W be a neighborhood in $\mathbb{R}^n$ of $x_0 \in \partial\Omega$ and let $V = W \cap \Omega$, we denote by $\mathscr{C}_0^\infty(\overline{V}) = \{u \in \mathscr{C}^\infty(V),$ such that there exists $v \in \mathscr{C}_0^\infty(W),$ with $u = v_{|V}\}$. For a function $w \in L^2(V)$ we denote the norm by $\|v\|_{L^2(V)}$ and for a fonction $v \in L^2(\partial V)$ we denote the norm by $|v|_{L^2(\partial V)}$.

Theorem 3.1. *Let $x_0 \in \partial\Omega$ and P be an elliptic operator of order 2 with real coefficients and $\phi \in \mathscr{C}^\infty$ satisfying* (3.1) *at x_0 and $d\phi(x_0) \neq 0$. Then there exist W a neighborhood of x_0 in $\mathbb{R}^n$, $C > 0$ and $\tau_0 > 0$ such that for all $u \in \mathscr{C}_0^\infty(\overline{V})$, where $V = W \cap \Omega$, and all $\tau \geq \tau_0$,*

$$\begin{aligned}&\sum_{|\alpha|\leq 1} \tau^{3-2|\alpha|} \|e^{\tau\phi(x)} D^\alpha u(x)\|^2_{L^2(V)}\\ \leq &C\|e^{\tau\phi(x)} Pu(x)\|^2_{L^2(V)} + \sum_{|\alpha|\leq 1} \tau^{3-|\alpha|} |(e^{\tau\phi(x)} D^\alpha u)_{|\partial\Omega}|^2_{L^2(\partial V)},\end{aligned} \tag{3.2}$$

where $\partial V = \overline{V} \cap \partial\Omega$.

Theorem 3.2. *Let $x_0 \in \partial\Omega$ and P be an elliptic operator of order 2 with real coefficients and $\phi \in \mathscr{C}^\infty$ satisfying* (3.1) *at x_0 and $\partial_\nu \phi(x_0) < 0$, where ∂_ν is the exterior normal*

derivative at ∂V. Then there exist W a neighborhood of x_0 in $\mathbb{R}^n$, $C>0$ and $\tau_0>0$ such that for all $u\in \mathscr{C}_0^\infty(\overline{V})$, where $V=W\cap\Omega$, satisfying $u_{|\partial V}=0$ and all $\tau\geq\tau_0$,

$$\begin{aligned}&\sum_{|\alpha|\leq 1}\tau^{3-2|\alpha|}\|e^{\tau\phi(x)}D^\alpha u(x)\|^2_{L^2(V)}+\tau|(e^{\tau\phi(x)}\partial_\nu u)_{|\partial\Omega}|^2_{L^2(\partial V)}\\ \leq& C\|e^{\tau\phi(x)}Pu(x)\|^2_{L^2(V)}.\end{aligned}\tag{3.3}$$

We can find the proofs of Theorem 3.1 and 3.2 in [91].

Remark 3.3. We can obtain a Carleman estimate without the boundary condition $u_{|\partial V}=0$ but in this case we must add at the right hand side of (3.3) the terms $\tau^3|(e^{\tau\phi(x)}u)_{|\partial\Omega}|^2_{L^2(\partial V)}$ and $\tau|(e^{\tau\phi(x)}X_ju)_{|\partial\Omega}|^2_{L^2(\partial V)}$, where $(X_j)_j$ is a basis of vector fields tangent to ∂V.

We have the same result for the Neumann boundary condition.

Theorem 3.4. *Let $x_0\in\partial\Omega$ and P be an elliptic operator of order 2 with real coefficients and $\phi\in\mathscr{C}^\infty$ satisfying* (3.1) *at x_0 and $\partial_\nu\phi(x_0)<0$, where ∂_ν is the exterior normal derivative at ∂V. Then there exist W a neighborhood of x_0 in $\mathbb{R}^n$, $C>0$ and $\tau_0>0$ such that for all $u\in\mathscr{C}_0^\infty(\overline{V})$, where $V=W\cap\Omega$, satisfying $\partial_\nu u_{|\partial V}=0$ and all $\tau\geq\tau_0$,*

$$\begin{aligned}&\sum_{|\alpha|\leq 1}\tau^{3-2|\alpha|}\|e^{\tau\phi(x)}D^\alpha u(x)\|^2_{L^2(V)}+\sum_{|\alpha|\leq 1}\tau^{3-|\alpha|}|(e^{\tau\phi(x)}D^\alpha u)_{|\partial\Omega}|^2_{L^2(\partial V)}\\ \leq& C\|e^{\tau\phi(x)}Pu(x)\|^2_{L^2(V)}.\end{aligned}\tag{3.4}$$

Remark 3.5. Here also we can obtain a Carleman estimate without condition $\partial_\nu u_{|\partial V}=0$ if we add the term $\tau|(e^{\tau\phi(x)}\partial_\nu u)_{|\partial\Omega}|^2_{L^2(\partial V)}$ at the right hand side of (3.4).

Remark 3.6. With the boundary terms at the right hand side (see Remaks 3.3 and 3.5), Theorems 3.2 and 3.4 imply Theorem 3.1. So Theorem 3.1 is relevant only locally in a neighborhood where $\partial_\nu\phi(x_0)\geq 0$. In [91] the proof is given for $\partial_\nu\phi(x_0)\neq 0$ but we can also prove the result for $\partial_\nu\phi(x_0)=0$.

Remark 3.7. We can find functions ϕ satisfying the assumption of previous theorems. Let φ such that $d\varphi(x_0)\neq 0$ then $\phi(x)=e^{\lambda\varphi(x)}$ satisfies (3.1) at x_0 for λ large enough.

Following these three previous theorems we can give global Carleman estimates for global weight function ϕ. First for Dirichlet boundary condition and second for Neumann boundary condition.

Theorem 3.8. *Let ω be an open subset of Ω eventually empty. Assume that for all $x_0\in\overline{\Omega}\backslash\omega$ the condition* (3.1) *is fulfilled. Let $\Gamma\subset\partial\Omega$ a neighborhood of $\{x\in\partial\Omega,\ \partial_\nu\phi(x)\geq 0\}$. Then there exist $C>0$ and $\tau_0>0$ such that for all $u\in\mathscr{C}^\infty(\Omega)$ satisfying either $u_{|\partial\Omega}=0$ or $\partial_\nu u_{|\partial\Omega}=0$, and all $\tau>\tau_0$,*

$$\begin{aligned}&\sum_{|\alpha|\leq 1}\tau^{3-2|\alpha|}\|e^{\tau\phi(x)}D^\alpha u(x)\|^2_{L^2(\Omega)}+\sum_{|\alpha|\leq 1}\tau^{3-|\alpha|}|(e^{\tau\phi(x)}D^\alpha u)_{|\partial\Omega}|^2_{L^2(\partial\Omega)}\\ \leq& C\|e^{\tau\phi(x)}Pu(x)\|^2_{L^2(\Omega)}+\sum_{|\alpha|\leq 1}\tau^{3-2|\alpha|}\|e^{\tau\phi(x)}D^\alpha u(x)\|^2_{L^2(\omega)}\\ &+\sum_{|\alpha|\leq 1}\tau^{3-|\alpha|}|(e^{\tau\phi(x)}D^\alpha u)_{|\partial\Omega}|^2_{L^2(\Gamma)}.\end{aligned}\tag{3.5}$$

Remark 3.9. For geometric reasons it is not alway possible to have a weight ϕ satisfying (3.1) in $\overline{\Omega}$ or $\partial_\nu \phi(x) < 0$ for all $x \in \partial\Omega$. This is the reason for introducing the sets ω and Γ.

Remark 3.10. The proof of Theorem 3.8 follow the strategy developed by Hörmander [64, Lemma 8.3.1], which proves that Carleman estimate is a local property. If the estimate (3.5) is true locally in a neighborhood of all point of Ω we can gather all in a global estimate.

In the following propositions, we give a way to construct ϕ satisfying assumptions of Theorem 3.8.

Proposition 3.11. *Let ω be an open subset of Ω there exists $\varphi \in \mathscr{C}^\infty(\overline{\Omega})$ such that*

$$\begin{aligned}
&\varphi(x) = 0 \text{ for } x \in \partial\Omega \\
&\partial_\nu \varphi(x) < 0 \text{ for } x \in \partial\Omega \\
&d\varphi(x) \neq 0 \text{ for } x \in \Omega \setminus \omega.
\end{aligned}$$

Proposition 3.12. *Let Γ be an open subset of $\partial\Omega$ there exists $\varphi \in \mathscr{C}^\infty(\overline{\Omega})$ such that*

$$\begin{aligned}
&\varphi(x) = 0 \text{ for } x \in \partial\Omega \setminus \Gamma \\
&\partial_\nu \varphi(x) < 0 \text{ for } x \in \partial\Omega \setminus \Gamma \\
&d\varphi(x) \neq 0 \text{ for } x \in \Omega.
\end{aligned}$$

The proofs of both propositions can be found in [60]. The idea is to start with a morse function. Next we move the critical points, in ω, to prove Proposition 3.11 or outside Ω through Γ to prove Proposition 3.12, by a diffeomorphism constructed as the flow of a vector field.

Now we can verify that $\phi(x) = e^{\lambda\varphi(x)}$ satisfies the assumptions of Theorem 3.8 for λ large enough. We obtain the following theorem.

Theorem 3.13. *Let φ be the function constructed in Proposition 3.11 and $\phi(x) = e^{\lambda\varphi(x)}$ where λ is large enough such that condition* (3.1) *is satisfied. Then there exist $C > 0$ and $\tau_0 > 0$ such that for all $u \in \mathscr{C}^\infty(\Omega)$ satisfying either $u_{|\partial\Omega} = 0$ or $\partial_\nu u_{|\partial\Omega} = 0$, and all $\tau > \tau_0$,*

$$\begin{aligned}
&\sum_{|\alpha|\leq 1} \tau^{3-2|\alpha|} \|e^{\tau\phi(x)} D^\alpha u(x)\|^2_{L^2(\Omega)} + \sum_{|\alpha|\leq 1} \tau^{3-|\alpha|} |(e^{\tau\phi(x)} D^\alpha u)_{|\partial\Omega}|^2_{L^2(\partial\Omega)} \\
&\quad \leq C \|e^{\tau\phi(x)} Pu(x)\|^2_{L^2(\Omega)} + \sum_{|\alpha|\leq 1} \tau^{3-2|\alpha|} \|e^{\tau\phi(x)} D^\alpha u(x)\|^2_{L^2(\omega)}. \qquad (3.6)
\end{aligned}$$

Theorem 3.14. *Let φ be the function constructed in Proposition 3.12 and $\phi(x) = e^{\lambda\varphi(x)}$ where λ is large enough such that condition* (3.1) *is satisfied. Then there exist $C > 0$ and $\tau_0 > 0$ such that for all $u \in \mathscr{C}^\infty(\Omega)$ satisfying either $u_{|\partial\Omega} = 0$ or $\partial_\nu u_{|\partial\Omega} = 0$, and all $\tau > \tau_0$,*

$$\begin{aligned}
&\sum_{|\alpha|\leq 1} \tau^{3-2|\alpha|} \|e^{\tau\phi(x)} D^\alpha u(x)\|^2_{L^2(\Omega)} + \sum_{|\alpha|\leq 1} \tau^{3-|\alpha|} |(e^{\tau\phi(x)} D^\alpha u)_{|\partial\Omega}|^2_{L^2(\partial\Omega)} \\
&\quad \leq C \|e^{\tau\phi(x)} Pu(x)\|^2_{L^2(\Omega)} + \sum_{|\alpha|\leq 1} \tau^{3-|\alpha|} |(e^{\tau\phi(x)} D^\alpha u)_{|\partial\Omega}|^2_{L^2(\Gamma)}. \qquad (3.7)
\end{aligned}$$

Remark 3.15. There are a lot of Carleman estimate up the boundary proven in literature for different boundary conditions or different assumptions on coefficient regularity, see [56, 57, 59, 70, 99].

4. Applications, control, stabilization, related fields

Since Lions [98], there are lot of works on control, stabilization for partial differential equations. One way to study these problems is the microlocal analysis following Lebeau and al [17] and the microlocal defect measure, see [30, 32, 33, 62, 63, 103, 134] which concern wave equations. Another way is to use the return method see Coron [41] or spectral method based on Ingham inequality (see [81] a recent paper on the subject). We can see the book [138] for an introduction and results on the subject. Here we describe only some results obtained by Carleman estimates.

4.1. Approximate controllability for wave equation. Exact control for wave equation is well understood and requires the geometric control condition see [17]. The HUM (Hilbert Uniqueness Method) given by Lions [98] allows to relate approximate control to an uniqueness result. The statement is the following.

Fix $T > 0$ and Γ an open in $\partial\Omega$ where Ω satisfied assumption made in Section 3. Let $g \in L^2(\mathbb{R} \times \partial\Omega)$ supported in $(0,T) \times \Gamma$ and $v \in \mathscr{C}(\mathbb{R}, L^2(\Omega)) \cap \mathscr{C}^1(\mathbb{R}, H^{-1}(\Omega))$ be the solution of

$$\begin{cases} (\partial_{tt} + P)v = 0 \\ (v(0,\cdot), \partial_t v(0,\cdot)) = (0,0) \\ v_{|\partial\Omega} = g. \end{cases} \tag{4.1}$$

We denote by $S(g) = (v(T,\cdot), \partial_t v(T,\cdot)) \in L^2(\Omega) \oplus H^{-1}(\Omega)$. Let

$$\mathcal{F} = \{w \in L^2(\Omega) \oplus H^{-1}(\Omega), \exists g \in L^2(\mathbb{R}\times\partial\Omega) \text{ supported in } (0,T)\times\Gamma \text{ such that } S(g) = w\}.$$

We associate the following adjoint problem. Let $(u_0, u_1) \in H_0^1(\Omega) \oplus L^2(\Omega)$ and $u \in \mathscr{C}(\mathbb{R}, H_0^1(\Omega)) \cap \mathscr{C}^1(\mathbb{R}, L^2(\Omega))$ be the solution of

$$\begin{cases} (\partial_{tt} + P)u = 0 \\ (u(0,\cdot), \partial_t u(0,\cdot)) = (u_0, u_1) \\ u_{|\partial\Omega} = 0. \end{cases} \tag{4.2}$$

We denote by $K(u_0, u_1) = (\partial_\nu u)_{|(0,T)\times\Gamma}$. Of course problems (4.1) and (4.2) are well-posed in the spaces given. If $\mathcal{F} = L^2(\Omega) \oplus H^{-1}(\Omega)$ we say that the problem is exactly controllable. If $\mathcal{F}$ is dense in $L^2(\Omega) \oplus H^{-1}(\Omega)$ we say that the problem is approximate controllable.

Theorem 4.1 ([98, Chapter 2]). *The problem* (4.1) *is approximate controllable if and only if the following property is satisfied*

$$K(u_0, u_1) = 0 \Rightarrow (u_0, u_1) = (0,0) \tag{4.3}$$

If P has analytic coefficients then Property (4.3) can be proven by Holmgren theorem if T large enough. If P has $\mathscr{C}^\infty$ coefficients we cannot apply the uniqueness theorem 2.13

because the surfaces needs to prove that $u = 0$ are not pseudo-convex. In [120] we gave an uniqueness result adapted to this case, see also [67], [135], [68], [122], and [137] where analyticity with respect the "t" variable is a crucial assumption. In the following theorem we are more precise than a uniqueness theorem and we estimate the solution by the boundary data.

Theorem 4.2. *For Γ an open subset of $\partial\Omega$ and all $\beta \in (0,1)$, there exist constants $T > 0$ and $C > 0$ such that the solution of* (4.2) *where $(u_0, u_1) \in H_0^1(\Omega) \oplus L^2(\Omega)$, satisfies*

$$\|(u_0,u_1)\|_{L^2(\Omega)\oplus H^{-1}(\Omega)} \leq \frac{C\|(u_0,u_1)\|_{H^1(\Omega)\oplus L^2(\Omega)}}{\left(\log\left(2+\frac{\|(u_0,u_1)\|_{H^1(\Omega)\oplus L^2(\Omega)}}{\|K(u_0,u_1)\|_{L^2((0,T)\times\Gamma)}}\right)\right)^{\beta}}.$$

If $K(u_0, u_1) = 0$ the denominator in the estimate above is ∞ then $(u_0, u_1) = (0, 0)$. We find the uniqueness result. Actually T found by the proof of Theorem 4.2 is larger than the one found in the uniqueness result and, in particular, T is not optimal.

Remark 4.3. This theorem is proven by Phung [113], in [121] we proved this result with $\beta = 1/2$. This kind of estimates was proven by Johns [76] (with different norms that those used here) in the context of Holmgren theorem.

By duality we can estimate the cost function. We define a norm on $\mathcal{F}$ by

$$\|v\|_{\mathcal{F}} = \inf\{\|g\|_{L^2((0,T)\times\Gamma)}, \text{ such that } g \text{ is supported in } (0,T)\times\Gamma \text{ and satisfies } Sg = v\}.$$

Of course $\mathcal{F}$ and the norm depend of T and Γ. In general the g's such that $Sg = v$ are not unique.

Theorem 4.4. *For all $\alpha > 1$, there exists $C > 0$ such that for all $v = (v_0, v_1) \in H_0^1(\Omega) \oplus L^2(\Omega)$, such that $\|(v_0, v_1)\|_{H^1(\Omega)\oplus L^2(\Omega)} \leq 1$ and all $\varepsilon > 0$ there exist $w = (w_0, w_1) \in \mathcal{F}$ and $u = (u_0, u_1) \in L^2(\Omega) \oplus H^{-1}(\Omega)$ such that $v = u + w$ where $\|u\|_{L^2(\Omega)\oplus H^{-1}(\Omega)} \leq \varepsilon$ and $\|w\|_{\mathcal{F}} \leq Ce^{C/\varepsilon^{\alpha}}$.*

Remark 4.5. In [120] we proved this result with $\alpha = 2$ but using the estimate given in Theorem 4.2, we can prove Theorem 4.4. If P has analytic coefficient, there is, in [89], a more precise result on the space $\mathcal{F}$, in particular a consequence of this result is that we can take $\alpha = 1$.

4.2. Null control for heat equation.

4.2.1. Spectral approach. This method is initiated by Fattorini and Russell in [52, 123] and it is used in [75, 91, 93]. It follows several steps, first the Carleman estimate allows to prove an interpolation estimate (see Theorem 4.6), second we deduce an estimate on the sum on eigenfunctions (see Theorem 4.8), third this gives a control for the low frequency with a control on the cost, fourth by alternatively applying control and dissipation we prove the null control (see Theorem 4.10).

We use the notations given in the begin of section 3. We introduce an other variable denoted by s and by $Q = D_s^2 + P$. Let $S > 0$, $\omega \subset \{0\} \times \Omega$ be an open, $X = (0, 3S) \times \Omega$ and $Y = (S, 2S) \times \Omega$.

Theorem 4.6. *There exist $C > 0$ and $\delta \in (0,1)$ such that for all $v \in \mathscr{C}^\infty(\overline{X})$ satisfying $v = 0$ on $(0,4S) \times \partial\Omega$ and on $\{0\} \times \Omega$, we have*

$$\|v\|_{H^1(Y)} \leq C \left(\|Qv\|_{L^2(X)} + |\partial_\nu v_{|\omega}|_{L^2(\omega)} \right)^\delta \|v\|^{1-\delta}_{H^1(X)}. \tag{4.4}$$

Remark 4.7. This kind of estimate was first proven, using Carleman estimate, in [14, 73]. Carleman estimate allows to prove local estimate analogous to (4.4). Next we can patch together these estimates to prove (4.4). This estimate is useful below but we can prove by the same Carleman estimates other interpolation estimate useful in other context. For instance, we can change the boundary conditions on v, or the control term $\partial_\nu v_{|\omega}$ (see [40, 77, 112]).

In the following theorems, P, with the Dirichlet boundary condition, is assumed self-adjoint, positive. We denote by φ_j the normalized eigenfunctions of P , and λ_j the associated eigenvalues.

Theorem 4.8. *There exist $C > 0$ such that for all μ and $u = \sum_{\lambda_j \leq \mu^2} a_j \varphi_j$, we have*

$$\|u\|_{L^2(\Omega)} \leq C e^{C\mu} \|u\|_{L^2(\omega)}. \tag{4.5}$$

Remark 4.9. This estimate is optimal for sum of eigenfunction (see [75] and [88]). There exist examples proving that the estimate (4.5) is optimal yet for a sequence of eigenfunctions (see [22, 111]) but it is not optimal, for a sequence of eigenfunctions, for all domain as we can see on a cube in $\mathbb{R}^n$.

To prove Theorem 4.8 we define $v(s,\cdot) = \sum_{\lambda_j \leq \mu^2} (a_j/\sqrt{\lambda_j}) \sinh(\sqrt{\lambda_j} s) \varphi_j$. It is easy to see that $Qv = 0$ and $\partial_\nu v = -\partial_s v = -\sum_{\lambda_j \leq \mu^2} a_j \varphi_j$ on $s = 0$. The estimations by below or by above of $\|v\|_{H^1(Y)}$ and $\|v\|_{H^1(X)}$ and Theorem 4.6, give the factor $e^{C\mu}$.

Theorem 4.10. *Let $T > 0$ and ω be an open in Ω. There exists $C > 0$, such that for all $u_0 \in L^2(\Omega)$, there exist $g \in L^2((0,T) \times \Omega)$ supported in $(0,T) \times \omega$ such that the solution u of*

$$\begin{cases} \partial_t u + Pu = g \text{ in } (0,T) \times \Omega) \\ u_{|(0,T)\times\partial\Omega} = 0 \\ u(0,\cdot) = u_0, \end{cases} \tag{4.6}$$

satisfies $u(T,\cdot) = 0$. Moreover we can choose g such that $\|g\|_{L^2((0,T)\times\Omega)} \leq C \|u_0\|_{L^2(\Omega)}$.

Remark 4.11. The problem (4.6) is well-posed. The function g is called the control of u_0. It is easy to deduce a control to the trajectory namely if v is a solution to (4.6) with $g = 0$ we can find for all u_0 a function g such that the solution of (4.6) satisfy $u(T,\cdot) = v(T,\cdot)$. In [114] we can find estimate of control in spirit of section 4.1. This result does not use Carleman estimate. We can find also, in [55, 104] estimates on the cost of control in small time. This cost is estimate by $Ce^{C/T}$ when T is small. Actually we can prove that the attainable set is larger than the trajectory and contains $e^{-TP^{1/2}} u_0$ for all $u_0 \in L^2(\Omega)$ (see [107]).

Remark 4.12. The proof is based first on a control for the first eigenfunctions. We can prove following Theorem 4.8 that we can find a control g such that $u(T,\cdot)$ is orthogonal to φ_j for all j satisfying $\lambda_j \leq \mu^2$ and g satisfies $\|g\|_{L^2(0,T)\times\Omega)} \leq Ce^{K\mu}$. Letting the solution dissipate for $t > T$ the L^2 norm of the solution decrease at less of $e^{-c\mu^2}$. As we

dissipate more than energy given to the system, the final solution after application of control and dissipation is smaller than the initial solution. Repeating the procedure on an infinite numbers of interval the final solution is null. Of course we must more precise because the K and c given above depend on the interval of time T and this time goes to 0. But we can estimate precisely K and c with respect T and handle the procedure. This strategy works in other contexts, see [1, 85, 87, 105–108].

4.2.2. Fursikov-Imanuvilov method. Carleman estimate are proven in parabolic context to deduce unique continuation (see [109, 127, 131]) but we cannot use these estimates to prove observability estimates because the domain in time of the estimated quantity is smaller than the observed domain. This phenomena is analogous to (4.4) where Y is smaller than X. By this method we can only prove uniqueness result and approximative controllability. To remedy it, Fursikov and Imanuvilov introduced singular weights. Consequence the cut-off function in time is in the weight. This is possible because the order of the operator is one in time variable and is two in spatial variables.

Theorem 4.13. *Let $T > 0$, there exist $C > 0$, $\lambda > 0$ and $\tau_0 > 0$ such that for all $u \in \mathscr{C}^\infty((0,T)\times\Omega)$, satisfying $u(t,x) = 0$ for $(t,x) \in (0,T)\times\partial\Omega$, we have for all $\tau > \tau_0$,*

$$\tau^3\|\eta^{3/2}e^{\tau\eta\phi}u\|^2_{L^2((0,T)\times\Omega)}+\tau\|\eta^{1/2}e^{\tau\eta\phi}\nabla_x u\|^2_{L^2((0,T)\times\Omega)} \leq C\|e^{\tau\eta\phi}(\partial_t + P)u\|^2_{L^2((0,T)\times\Omega)} + C\|e^{\tau\eta\phi}u\|^2_{L^2((0,T)\times\omega)}. \tag{4.7}$$

where $\eta(t) = \frac{T^2}{t(T-t)}$, $\phi(x) = e^{\lambda\varphi(x)} - e^{\lambda K}$ with φ defined in Proposition 3.11 and $K \geq 2\max_{x\in\Omega}|\varphi(x)|$.

Remark 4.14. In the left hand side of (4.7), we can add first derivative with respect t and Pu with a factor $(\tau\eta)^{-1/2}$ in the norm, see the original proof given by Fursikov and Imanuvilov [60]. This estimate can be proven for operator P with coefficients depending of t. This kind of estimate can be use for nonlinear equations, for heat equations with non global lipschitz non linearities (see [45, 55]) or for Navier-Stokes and Boussinesq systems (see [48, 53, 54, 61]). Same kind of estimates can be proven for other boundary conditions or transmission conditions, see [27, 46, 71, 84, 86]. In [47, 139] there are same kind of Carleman estimates for heat equations with a potential in $|x|^{-2}$. All these results show that the method is very flexible.

Fixing τ in (4.7) and using the fact that the parabolic problem is well-posed we can obtain the observability estimate

Theorem 4.15. *Let $u \in \mathscr{C}([0,T], L^2(\Omega))$ be the solution of the problem* (4.6)*, there exists $C > 0$ such that*

$$\|u(T,\cdot)\|_{L^2(\Omega)} \leq C\|u(t,x)\|_{L^2((0,T)\times\omega)}.$$

Remark 4.16. This estimate is equivalent to the null controllability. This is another way to prove Theorem 4.10.

Remark 4.17. There is some results for parabolic systems using Carleman estimates. The goal is to find Kalman type condition to control. Typically when the number of control is smaller than the size of system. We can see [7, 8, 78] and references of these papers for results in this direction.

4.3. Stabilization for wave equation. In this section we study the decay of the energy for the wave equation with a damping which is localized either in a subset of domain Ω or on a subset of the boundary Γ.

We consider the following problem, let $\omega \subset \Omega$ be an open and $a \in \mathscr{C}_0^\infty(\omega)$ where $a(x) \geq 0$ for all $x \in \omega$. Let $(u_0, u_1) \in H_0^1(\Omega) \oplus L^2(\Omega)$, and u be the solution of

$$\left\{ \begin{array}{l} \partial_{tt}^2 u(t,x) - \Delta u(t,x) + a(x)\partial_t u(t,x) = 0 \text{ in } (0,+\infty)\times\Omega, \\ (u(0,\cdot), \partial_t u(0,\cdot) = (u_0, u_1), \\ u_{|(0,+\infty)\times\partial\Omega} = 0. \end{array} \right. \tag{4.8}$$

The problem is well-posed and for simplicity we consider Δ but the results are true for all elliptic operator of order two, self-adjoint. We define the energy by $E(u)(t) = \int_\Omega \left(|\nabla u(t,x)|^2 + |\partial_t u(t,x)|^2\right) dx$. If $a \equiv 0$ then $E(u)(t) = E(u)(0)$, and in general

$$E(u)(T) - E(u)(0) = -2\int_0^T \int_\omega a(x)|\partial_t u(t,x)|^2 dxdt.$$

As $a \geq 0$, E(u)(t) decreases and we can prove that $E(u)(t) \to 0$ when $t \to +\infty$. The goal is to quantify the decay. Under Geometric Control Condition it is proven that there exists $C > 0$ such that $E(u)(t) \leq C\|(u_0,u_1)\|^2_{H^1(\Omega)\oplus L^2(\Omega)} e^{-t/C}$ (see [17], [90], [128], [9]). The interest of the following theorem is when the Geometric Control Condition is not satisfied. We define

$$A = \begin{pmatrix} 0 & Id \\ \Delta & -a(x) \end{pmatrix},$$

the generator of the semi-group associated with the problem (4.8).

Theorem 4.18. *Assume that $a \not\equiv 0$ then for all $k > 0$, there exist $C > 0$, such that for all $(u_0, u_1) \in \mathcal{D}(A^k)$,*

$$(E(u)(t))^{1/2} \leq \frac{C\|(u_0,u_1)\|_{\mathcal{D}(A^k)}}{\log(2+t)^k}.$$

Actually this result is the consequence of an estimate on the resolvent.

Theorem 4.19. *There exist $C > 0$ and $\delta > 0$ such that for all $\lambda \in \mathbb{R}$,*

$$\|(i\lambda - A)^{-1}\| \leq Ce^{\delta|\lambda|},$$

where $\|\cdot\|$ is the norm operator between $L^2(\Omega) \to L^2(\Omega)$.

Remark 4.20. The proof of Theorem 4.19 is in [90]. Theorem 4.8 was proven with a $\log\log t$ in numerator of the right hand side and improved in [31]. See also [18] for a simpler proof of the link between the estimate on resolvent and the energy decay. We can find also similarly results in [58] where the regularity assumptions are weaker.

We have the same kind of result for the boundary damping.

We consider the following problem, let $\Gamma \subset \partial\Omega$ be an open of $\partial\Omega$ and $a \in \mathscr{C}_0^\infty(\Gamma)$ where $a(x) \geq 0$ for all $x \in \partial\omega$. We define

$$A = \begin{pmatrix} 0 & Id \\ \Delta & 0 \end{pmatrix},$$

and $\mathcal{D}(A) = \{(u_0, u_1) \in H^1(\Omega) \oplus L^2(\Omega),\ A(u_0, u_1) \in H^1(\Omega) \oplus L^2(\Omega),\ (\partial_\nu u_0 + au_1)_{|\partial\Omega} = 0\}$.

Let $(u_0, u_1) \in H_0^1(\Omega) \oplus L^2(\Omega)$, and u be the solution of

$$\begin{cases} \partial_{tt}^2 u(t,x) - \Delta u(t,x) = 0 \text{ in } (0,+\infty) \times \Omega, \\ (u(0,\cdot), \partial_t u(0,\cdot) = (u_0, u_1), \\ \partial_\nu u + a(x)\partial_t u(t,x) = 0 \text{ on } (0,+\infty) \times \partial\Omega. \end{cases} \tag{4.9}$$

The problem is well-posed, in the space given, by classical Hille-Yosida theorem. The energy decreases and we have $E(u)(T) - E(u)(0) = -2\int_0^T \int_\Gamma a(x)|\partial_t u(t,x)|^2 d\sigma dt$, where $d\sigma$ is the measure on the boundary.

Theorem 4.21. *Assume that $a \not\equiv 0$ then for all $k > 0$, there exists $C > 0$, such that for all $(u_0, u_1) \in \mathcal{D}(A^k)$,*

$$(E(u)(t))^{1/2} \leq \frac{C\|(u_0, u_1)\|_{\mathcal{D}(A^k)}}{\log(2+t)^k}.$$

This result follows from the estimate on the resolvent.

Theorem 4.22. *There exist $C > 0$ and $\delta > 0$ such that for all $\lambda \in \mathbb{R}$,*

$$\|(i\lambda - A)^{-1}\| \leq Ce^{\delta|\lambda|},$$

where $\|\cdot\|$ is the norm operator between $L^2(\Omega) \to L^2(\Omega)$.

Remark 4.23. Theorem 4.22 is proven in [92] and Theorem 4.21 follows from Theorem 4.19. Other results for this kind of problems was proven in [25, 40, 57, 77, 112].

4.4. Local energy decay for wave equation. Local energy for the wave equation decreases exponentially under non trapping condition in odd dimension and polynomial in even dimension, see [82, 100, 110]. Here we give the Burq result [31] proven without any geometrical conditions but the decay is only logarithmic.

Let K be a compact set in $\mathbb{R}^n$, such that $\mathring{K}$ is a open set with $\mathscr{C}^\infty$ boundary. Let $\Omega = \mathbb{R}^n \setminus K$. We assume P elliptic, self-adjoint and there exists $R > 0$ such that $P = \Delta$ for $|x| > R$. It is convenient to introduce the semi-group generator B of the wave equation given by

$$B = \begin{pmatrix} 0 & Id \\ -P & 0 \end{pmatrix},$$

where P is the operator given by P in Ω with Dirichlet boundary condition. Let U be the solution of $\partial_t U - BU = 0$ with $U(0) = (u_1^0, u_2^0) \in H = H_0^1(\Omega) \oplus L^2(\Omega)$. This solution will be denoted by $U(t) = e^{tB}(u_1^0, u_2^0)$. Let $U(t) = (u_1(t), u_2(t))$ we have $\partial_{tt}^2 u_1 - Pu_1 = 0$. Let $R_0 > 0$, the local energy is given by $E_{\text{loc}}(U,t) = (1/2)\int_{\Omega \cap B(0,R_0)} (|\nabla u_1(t,x)|^2 + |\partial_t u_1(t,x)|^2)dx$ where ∇ is associated with the Riemannian metric defining P.

Theorem 4.24. *Let $k > 0$ and $R_1 > 0$, there exists $C > 0$ such that for all $(u_1^0, u_2^0) \in \mathcal{D}(B^k)$ supported in $B(0, R_1)$, the solution $U(t) = e^{tB}(u_1^0, u_2^0)$ satisfies*

$$E_{\text{loc}}(U,t)^{1/2} \leq \frac{C\|U(0)\|_{\mathcal{D}(B^k)}}{\log(2+t)^k}.$$

As for Theorem 4.18 the proof is the consequence of a stationary result on the resolvent. We define $R(\tau)$ such that $(P-\lambda^2)R(\lambda)f = f$. If $\operatorname{Im}\lambda < 0$, $R(\lambda)$ is well defined and we can extend $R(\lambda)$ in $\mathbb{C}\setminus\{0\}$ as a meromorphic operator from $L^2_{\rm comp}$ to $H^1_{0,\rm loc}(\Omega)$. The resolvent $(i\lambda - B)^{-1}$ is related with $R(\lambda)$ by the formula

$$(i\lambda - B)^{-1} = \begin{pmatrix} i\lambda R(\lambda) & R(\lambda) \\ -I - \lambda^2 R(\lambda) & i\lambda R(\lambda) \end{pmatrix}. \tag{4.10}$$

Theorem 4.25. *Let χ_1 and χ_2 be $\mathscr{C}_0^\infty(\mathbb{R}^n)$ functions. There exist $C_0 > 0$ and $C_1 > 0$ such that for all $\lambda \in \mathbb{R}$ we have*

$$\|\chi_1 R(\lambda)\chi_2\|_{\mathcal{L}(L^2(\Omega),H^1_0(\Omega))} \leq C_0 e^{C_1\lambda}.$$

Theorem 4.24 is the consequence of (4.10) and Theorems 4.22 and 4.25.

Remark 4.26. The same kind of result can be proven for elasticity system (see [24]). In this case, even for convex obstacle, the decay of local energy cannot be better than the inverse of every power of t (see [133]).

4.5. Other problems related to Carleman estimate. Carleman estimates are tools used to the study of control for other equations and in other field related to control theory and stabilization.

The null-control for degenerate parabolic equations was studied in [35, 36], in [21] we can find results on Grushin operator and in [20, 22, 23] results on Kolmogorov operator. These problems are still open and require other researches.

The optimal control obtains on the discretized equation of a partial differential equation has oscillation in general even if the control for partial differential equation does not oscillate. This problem is not well understood and some approach uses Carleman estimate (see [28, 29, 49])

The stochastic equations was studied with Carleman estimate see ([142] for references and [97]).

Optimal control in time was studied under a bound constraint on the control. A bang-bang property can be proven and this requires to control heat equation on measurable set. Results are proven for this problem in [11, 115].

The Carleman estimates are tools used in the inverse problems. We can see [69] for a review in subject. It is impossible to give a complete bibliography on the subject here. See some recent works, [19, 26, 44, 72, 101], to find more references.

References

[1] G. Alessandrini, L. Escauriaza, *Null-controllability of one-dimensional parabolic equations*, ESAIM Control Optim. Calc. Var. 14 (2008), 284–293.

[2] S. Alinhac, *Non-unicité du problème de Cauchy*, Ann. of Math. 117 (1983), 77–108.

[3] S. Alinhac and M.S. Baouendi, *Construction de solutions nulles et singulières pour des opérateurs de type principal*, Séminaire Goulaouic-Schwartz (1978/1979) École Polytech. Palaiseau.

[4] _____, *Uniqueness for the characteristic Cauchy problem and strong unique continuation for higher order partial differential inequalities*, Amer. J. Math. **102** (1980), 179–217.

[5] _____, *A counterexample to strong uniqueness for partial differential equations of Schrödinger's type*, Comm. Partial Differential Equations **19** (1994), 1727–1733.

[6] _____, *A nonuniqueness result for operators of principal type*, Math. Z. **220** (1995), 561–568.

[7] F. Ammar-Khodja, A. Benabdallah, C. Dupaix, M. González-Burgos, *A Kalman rank condition for the localized distributed controllability of a class of linear parbolic systems*, J. Evol. Equ. **9** (2009), 267–291.

[8] _____, *A generalization of the Kalman rank condition for time-dependent coupled linear parabolic systems*, Differ. Equ. Appl. **1** (2009), 427–457.

[9] N. Anantharaman, *Spectral deviations for the damped wave equation*, Geom. Funct. Anal. **20** (2010), 593–626.

[10] D.D. Ang, M. Ikehata, D.D. Trong, and M. Yamamoto, *Unique continuation for a stationary isotropic Lamé system with variable coefficients*, Comm. Partial Differential Equations **23** (1998), 371–385.

[11] J. Apraiz, L. Escauriaza, G. Wang, and C. Zhang, *Observability Inequalities and Measurable Sets*, To appear in J. Eur. Math. Soc.

[12] N. Aronszajn, *A unique continuation theorem for solutions of elliptic partial differential equations or inequalities of second order*, J. Math. Pures Appl. **36** (1957), 235–249.

[13] H. Bahouri, *Non unicité du problème de Cauchy pour des opérateurs à symbole principal réel*, Comm. Partial Differential Equations **8** (1983), 1521–1547.

[14] H. Bahouri, *Dépendance non linéaire des données de Cauchy pour les solutions des équations aux dérivées partielles*, J. Math. Pures Appl. **66** (1987), 127–138.

[15] H. Bahouri and L. Robbiano, *Unicité de Cauchy pour des opérateurs faiblement hyperboliques*, Hokkaido Math. J. **16** (1987), 257–275.

[16] B. Barceló, C. E. Kenig, A. Ruiz, and C. D. Sogge, *Weighted Sobolev inequalities and unique continuation for the Laplacian plus lower order terms*, Illinois J. Math. **32** (1988), 230–245.

[17] C. Bardos, G. Lebeau., and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim. **30** (1992), 1024–1065.

[18] C.J.K Batty and T. Duyckaerts, *Non-uniform stability for bounded semi-groups on Banach spaces*, J. Evol. Equ. (2008), 765–780.

[19] L. Baudouin, É. Crépeau, and J. Valein, *Global Carleman estimate on a network for the wave equation and application to an inverse problem*, Math. Control Relat. Fields **1** (2011), 307–330.

[20] K. Beauchard, *Null controllability of Kolmogorov-type equations*, Mathematics of Control, Signals, and Systems **26** (2014), 145–176.

[21] K. Beauchard, P. Cannarsa, and R. Guglielmi, *Null controllability of Grushin-type operators in dimension two*, J. Eur. Math. Soc. **16** (2014), 67–101.

[22] K. Beauchard, B. Helffer, R. Henry, and L. Robbiano, *Degenerate parabolic operators of Kolmogorov type with a geometric control condition*, hal-00863056.

[23] K. Beauchard. and E. Zuazua, *Some controllability results for the 2D Kolmogorov equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire **26** (2009), 1793–1815.

[24] M. Bellassoued, *Distribution of resonances and decay rate of the local energy for the elastic wave equation*, Comm. Math. Phys. **215** (2000), 375–408.

[25] ———, *Decay of solutions of the elastic wave equation with a localized dissipation*, Ann. Fac. Sci. Toulouse Math. **12** (2003), 267–301.

[26] M. Bellassoued and M. Yamamoto, *Carleman estimate and inverse source problem for Biot's equations describing wave propagation in porous media*, Inverse Problems **29** (2013), 115002.

[27] A. Benabdallah, Y. Dermenjian, and J. Le Rousseau, *Carleman estimates for the one-dimensional heat equation with a discontinuous coefficient and applications to controllability and an inverse problem*, J. Math. Anal. Appl. **336** (2007), 865–887.

[28] F. Boyer, F. Hubert, and J. Le Rousseau, *Discrete Carleman estimates for elliptic operators and uniform controllability of semi-discretized parabolic equations*, J. Math. Pures Appl. **93** (2010) , 240–276.

[29] F. Boyer and J. Le Rousseau, *Carleman estimates for semi-discrete parabolic operators and application to the controllability of semi-linear semi-discrete parabolic equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire published on line.

[30] N. Burq, *Mesures semi-classiques et mesures de défaut, Séminaire Bourbaki*, Vol. 1996/97 Astérisque **245** (1997), 167–195.

[31] ———, *Décroissance de l'énergie locale de l'équation des ondes pour le problème extérieur et absence de résonance au voisinage du réel* Acta Math. (1998), 1–29.

[32] N. Burq and P. Gérard, *Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes*, C. R. Acad. Sci. Paris Sér. I Math. **325** (1997), 749–752.

[33] N. Burq and G. Lebeau, *Mesures de défaut de compacité, application au système de Lamé*, Ann. Sci. École Norm. Sup. **34** (2001), 817–870.

[34] A.P. Calderón, *Uniqueness in the Cauchy problem for partial differential equations*, Am. J. Math. **80** (1958), 16–36.

[35] P. Cannarsa, P. Martinez, and J. Vancostenoble, *Null controllability of degenerate heat equations*, Adv. Differential Equations **10** (2005), 153–190.

[36] P. Cannarsa, J. Tort, and M. Yamamoto, *Unique continuation and approximate controllability for a degenerate parabolic equation*, Appl. Anal. **91** (2012), 1409–1425.

[37] T. Carleman, *Sur un problème d'unicité pur les systèmes d'équations aux dérivées partielles à deux variables indépendantes*, Ark. Mat., Astr. Fys. **26** (1939) n° 17.

[38] S. Chanillo and E. Sawyer, *Unique continuation for $\Delta+v$ and the C*. Fefferman-Phong class Trans. Amer. Math. Soc. **318** (1990), 275–300.

[39] F. Colombini and C. Grammatico, *Some remarks on strong unique continuation for the Laplace operator and its powers*, Comm. Partial Differential Equations **24** (1999), 1079–1094.

[40] P. Cornilleau and L. Robbiano, *Carleman estimates for the Zaremba boundary condition and stabilization of waves*, Amer. J. Math. **136** (2014) 393–444.

[41] J.-M. Coron, *Control and nonlinearity*, Mathematical Surveys and Monographs 136 (2007) American Mathematical Society.

[42] B. Dehman and L. Robbiano, *La propriété du prolongement unique pour un système elliptique, Le système de Lamé*, J. Math. Pures Appl. **72** (1993), 475–492.

[43] D. Dos Santos Ferreira, *Sharp L^p Carleman estimates and unique continuation* Duke Math. J. **129** (2005), 503–550.

[44] D. Dos Santos Ferreira, C.E. Kenig, M. Salo, and G. Uhlmann, *Limiting Carleman weights and anisotropic inverse problems*, Invent. Math. **178** (2009), 119–171.

[45] A. Doubova, E. Fernandez-Cara, M. Gonzalez-Burgos, and E. Zuazua, *On the controllability of parabolic systems with a nonlinear term involving the state and the gradient*, SIAM J. Control Optim. **41** (2002), 798–819.

[46] A. Doubova, A. Osses, and J.-P. Puel, *Exact controllability to trajectories for semilinear heat equations with discontinuous diffusion coefficients*, ESAIM Control Optim. Calc. Var. **8** (2002), 621–661.

[47] S. Ervedoza, *Control and stabilization properties for a singular heat equation with an inverse-square potential*, Comm. Partial Differential Equations **33** (2008), 1996–2019.

[48] S. Ervedoza, O. Glass, S. Guerrero, and J.-P. Puel, *Local exact controllability for the one-dimensional compressible Navier-Stokes equation*, Arch. Ration. Mech. Anal. **206** (2012), 189–238.

[49] S. Ervedoza and F. de Gournay, *Uniform stability estimates for the discrete Calderón problems*, Inverse Problems **27** (2011) 125012.

[50] L. Escauruaza and L. Vega, *Carleman inequalities and the heat operator*, II Indiana Univ. Math. J. **50** (2001), 1149–1169.

[51] C. Fabre and G. Lebeau, *Prolongement unique des solutions de l'equation de Stokes* Comm. Partial Differential Equations **21** (1996), 573–596.

[52] H.O. Fattorini and D.L Russell, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal. **43** (1971), 272–292.

[53] E. Fernández-Cara, S. Guerrero, O. Yu. Imanuvilov, and J.-P. Puel, *Local exact controllability of the Navier-Stokes system*, J. Math. Pures Appl. **83** (2004), 1501–1542.

[54] _____, *Some controllability results for the N-dimensional Navier-Stokes and Boussinesq systems with $N-1$ scalar controls*, SIAM J. Control Optim. **45** (2006), 146–173.

[55] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: the linear case*, Adv. Differential Equations **5** (2000), 465–514.

[56] X. Fu, *Null controllability for the parabolic equation with a complex principal part*, J. Funct. Anal. **257** (2009), 1333–1354.

[57] _____, *Logarithmic decay of hyperbolic equations with arbitrary small boundary damping*, Comm. Partial Differential Equations **34** (2009), 957–975.

[58] _____, *Longtime behavior of the hyperbolic equations with an arbitrary internal damping*, Z. Angew. Math. Phys. **62** (2011), 667–680.

[59] X. Fu, J. Yong, and X. Zhang, *Exact controllability for multidimensional semilinear hyperbolic equations*, SIAM J. Control Optim. **46** (2007), 1578–1614.

[60] A. Fursikov and O. Yu. Imanuvilov, *Controllability of evolution equations*, Lecture Notes Series **34** (1996) Seoul National University, Korea.

[61] _____, *Exact controllability of the Navier-Stokes and Boussinesq equations*, Uspekhi Mat. Nauk **54** (1999), 93–146.

[62] P. Gérard, *Microlocal defect measures*, Comm. Partial Differential Equations **16** (1991), 1761–1794.

[63] P. Gérard and É. Leichtnam, *Ergodic properties of eigenfunctions for the Dirichlet problem*, Duke Math. J. **71** (1993), 559–607.

[64] L. Hörmander, *Linear partial differential operators*, Springer-Verlag, 1963.

[65] _____, *Non-uniqueness for the Cauchy problem*, (Colloq. Internat., Univ. Nice, Nice, 1974) Lecture Notes in Math., Vol 459 (1975), 36–72.

[66] _____, *The analysis of linear partial differential operators*, IV: Fourier integral operators. (1985) Springer-Verlag.

[67] _____, *A uniqueness theorem for second order hyperbolic differential equations*, Comm. Partial Differential Equations **17** (1992), 699–714.

[68] _____, *On the uniqueness of the Cauchy problem under partial analyticity assumptions*, Geometrical optics and related topics (Cortona, 1996) Progr. Nonlinear Differential Equations Appl. **32** (1997), 179–219

[69] O. Yu. Imanuvilov, *Controllability of evolution equations of fluid dynamics*, International Congress of Mathematicians Vol. III (2006), 1321–1338.

[70] O. Yu. Imanuvilov and J.-P. Puel, *Global Carleman estimates for weak solutions of elliptic nonhomogeneous Dirichlet problems*, Int. Math. Res. Not. **16** (2003), 883–913.

[71] O. Yu. Imanuvilov, J.-P. Puel, and M. Yamamoto, *Carleman estimates for parabolic equations with nonhomogeneous boundary conditions*, Chin. Ann. Math. Ser. B **30** (2009), 333–378.

[72] O. Yu. Imanuvilov and M. Yamamoto, *Uniqueness for inverse boundary value problems by Dirichlet-to-Neumann map on subboundaries*, Milan J. Math. **81** (2013), 187–258.

[73] V.M. Isakov, *On the uniqueness of the solution of the Cauchy problem*, Dokl. Akad. Nauk SSSR **255** (1980), 18–22.

[74] D. Jerison and C. Kenig, *Unique continuation and absence of positive eigenvalues for Schrödinger operators*, (appendix by E.M. Stein) Ann. of Math. **121** (1985), 463–494.

[75] D. Jerison and G. Lebeau, *Nodal sets of sums of eigenfunctions*, Chicago Lectures in Math. (1996), 223–239.

[76] F. John, *Continuous dependence on data for solutions of partial differential equations with a prescribed bound*, Comm. Pure Appl. Math. **13** (1960), 551–585.

[77] I. Kamoun Fathallah, *Logarithmic decay of the energy for an hyperbolic-parabolic coupled system*, ESAIM Control Optim. Calc. Var. **17** (2011), 801–835.

[78] O. Kavian and L. de Teresa, *Unique continuation principle for systems of parabolic equations*, ESAIM Control Optim. Calc. Var. **16** (2010), 247–274.

[79] C. Kenig, A. Ruiz, and C. D. Sogge, *Uniform Sobolev inequalities and unique continuation for second order constant coefficient differential operators*, Duke Math. J. **55** (1987), 329–347.

[80] H. Koch and D. Tataru, *Carleman estimates and unique continuation for second order parabolic equations with nonsmooth coefficients*, Comm. Partial Differential Equations **34** (2009), 305–366.

[81] V. Komornik and P. Loreti, *Discrete Ingham type inequalities and simultaneous observability of strings or beams*, J. Math. Anal. Appl. **351** (2009), 16–28.

[82] P.D. Lax and R.S. Phillips, *The acoustic equation with an indefinite energy form and the Schrödinger equation*, J. Functional Analysis **1** (1967), 37–83.

[83] P. Le Leborgne, *Unicité forte pour le produit de deux opérateurs, elliptiques d'ordre 2*, Indiana Univ. Math. J. **50** (2001), 353–381.

[84] J. Le Rousseau, *Carleman estimates and controllability results for the one-dimensional heat equation with BV coefficients*, J. Differential Equations **233** (2007), 417–447.

[85] J. Le Rousseau and L. Robbiano, *Carleman estimate for elliptic operators with coefficients with jumps at an interface in arbitrary dimension and application to the null controllability of linear parabolic equations*, Arch. Ration. Mech. Anal **195** (2010), 953–990.

[86] _____, *Local and global Carleman estimates for parabolic operators with coefficients with jumps at interfaces*, Invent. Math. **183** (2011), 245–336.

[87] J. Le Rousseau, M. Léautaud, and L. Robbiano, *Controllability of a parabolic system with a diffuse interface*, J. Eur. Math. Soc. (JEMS) **15** (2013), 1485–1574.

[88] G. Lebeau and J. Le Rousseau, *On Carleman estimates for elliptic and parabolic operators. Applications to unique continuation and control of parabolic equations*, ESAIM:COCV **18** (2012), 712–747.

[89] G. Lebeau, *Contrôle analytique. I. Estimations a priori*, Duke Math. J. **68** (1992), 1–30.

[90] _____, *Équation des ondes amorties. Algebraic and geometric methods in mathematical physics* , (Kaciveli, 1993) Math. Phys. Stud. **19** (1996), 73–109, Kluwer Acad. Publ.

[91] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations **20** (1995), 335–356.

[92] _____, *Stabilisation de l'équation des ondes par le bord*, Duke Math. J. **86** (1997), 465–491.

[93] G. Lebeau and E. Zuazua, *Null-controllability of a system of linear thermoelasticity*, Arch. Rational Mech. Anal. **141** (1998), 297–329.

[94] N. Lerner, *Unicité de Cauchy pour des opérateurs différentiels faiblement principalement normaux*, J. Math. Pures Appl. **64** (1985), 1–11.

[95] _____, *Carleman's and subelliptic estimates*, Duke Math. J. **56** (1988), 385–394.

[96] N. Lerner and L. Robbiano, *Unicité de Cauchy pour des opérateurs de type principal*, J. Analyse Math. **44** (1984/85), 32–66.

[97] H. Li and Q. Lü, *A quantitative boundary unique continuation for stochastic parabolic equations*, J. Math. Anal. Appl. **402** (2013), 518–526.

[98] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. Tome 1*, Recherches en Mathématiques Appliquées , vol. 8 Masson 1988.

[99] Q. Lü and Z. Yin, *The L^∞-null controllability of parabolic equation with equivalued surface boundary conditions*, Asymptot. Anal. **83** (2013), 355–378.

[100] R.B. Melrose and J. Sjöstrand, *Singularities of boundary value problems. I*, Comm. Pure Appl. Math. **31** (1978), 593–617.

[101] A. Mercado, A. Osses, and L. Rosier, *Inverse problems for the Schrödinger equation via Carleman inequalities with degenerate weights*, Inverse Problems **24**, (2008) 015017.

[102] K. Miller, *Nonunique continuation for uniformly parabolic and elliptic equations in self-adjoint divergence form with Hölder continuous coefficients*, Arch. Rational Mech. Anal. **54** (1974), 105–117.

[103] L. Miller, *Refraction of high-frequency waves density by sharp interfaces and semiclassical measures at the boundary*, J. Math. Pures Appl. **79** (2000), 227–269.

[104] _____, *Geometric bounds on the growth rate of null-controllability cost for the heat equation in small time*, J. Differential Equations **204** (2004), 202–226.

[105] _____, *On the null-controllability of the heat equation in unbounded domains*, Bull. Sci. Math. **129** (2005), 175–185.

[106] _____, *On the controllability of anomalous diffusions generated by the fractional Laplacian*, Math. Control Signals Systems **18** (2006), 260–271.

[107] _____, *On exponential observability estimates for the heat semigroup with explicit rates*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. **17** (2006), 351–366.

[108] _____, *A direct Lebeau-Robbiano strategy for the observability of heat-like semigroups*, Discrete Contin. Dyn. Syst. Ser. B **14** (2010), 1465–1485.

[109] S. Mizohata, *Unicité du prolongement des solutions pour quelques opérateurs différentiels paraboliques*, Mem. Coll. Sci. Univ. Kyoto. Ser. A. Math. **31** (1958), 219–239.

[110] C.S. Morawetz, *The decay of solutions of the exterior initial-boundary value problem for the wave equation*, Comm. Pure Appl. Math. **14** (1961), 561–568.

[111] B.-T. Nguyen and D.S. Grebenkov, *Localization of laplacian eigenfunctions in circular and elliptical domains*, SIAM J. Appl. Math. **73** (2013), 780–803.

[112] L. Ouksel, *Logarithmic stabilisation of a multidimensional structure by the boundary*, Asymptot. Anal. **80** (2012), 347–376.

[113] K.D. Phung, *Remarques sur l'observabilité pour l'équation de Laplace*, ESAIM Control Optim. Calc. Var. **9** (2003), 621–635.

[114] _____, *Note on the cost of the approximate controllability for the heat equation with potential*, J. Math. Anal. Appl. **295** (2004), 527–538.

[115] K.D. Phung, L. Wang, and C. Zhang, *Bang-bang property for time optimal control of semilinear heat equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire **31** (2014), 477–499.

[116] A. Pliś, *On non-uniqueness in Cauchy problem for an elliptic second order differential equation*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. **11** (1963), 95–100.

[117] R. Regbaoui, *Strong unique continuation results for differential inequalities*, J. Funct. Anal. **148** (1997), 508–523.

[118] _____, *Unique continuation for differential equations of Schrödinger's type*, Comm. Anal. Geom. **7** (1999), 303-323.

[119] L. Robbiano, *Sur les conditions de pseudo-convexité et l'unicité du problème de Cauchy*, Indiana Univ. Math. J. **36** (1987), 333–347.

[120] _____, *Théorème d'unicité adapté au contrôle des solutions des problèmes hyperboliques*, Comm. Partial Differential Equations **16** (1991), 789–800.

[121] _____, *Fonction de coût et contrôle des solutions des équations hyperboliques*, Asymptotic Anal. **10** (1995), 95–115.

[122] L. Robbiano and C. Zuily, *Uniqueness in the Cauchy problem for operators with partially holomorphic coefficients*, Invent. Math. **31** (1998), 493–539.

[123] D.L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math. **52** (1973), 189–211.

[124] X. Saint-Raymond, *L'unicité pour des problèmes de Cauchy caractéristiques*, Comm. Partial Differential Equations **7** (1982), 559–579.

[125] _____, *Non-unicité de Cauchy pour des opérateurs principalement normaux*, Indiana Univ. Math. J. **33** (1984), 847–858.

[126] _____, *Unicité de Cauchy à partir de surfaces faiblement pseudo-convexes*, Ann. Inst. Fourier (Grenoble) **39** (1989), 123–147.

[127] J.-C. Saut and B. Scheurer, *Unique continuation for some evolution equations*, J. Differential Equations **66** (1987), 118–139.

[128] J. Sjöstrand, *Asymptotic distribution of eigenfrequencies for damped wave equations*, Publ. Res. Inst. Math. Sci. **36** (2000), 573–611.

[129] C. D. Sogge, *Oscillatory integrals and unique continuation for second order elliptic differential equations*, J. Amer. Math. Soc. **2** (1989), 491–515.

[130] _____, *Strong uniqueness theorems for second order elliptic differential equations*, Amer. J. Math. **112** (1990), 943–984.

[131] _____, *A unique continuation theorem for second order parabolic differential operators*, Ark. Mat. **28** (1990), 159–182.

[132] _____, *Uniqueness in Cauchy problems for hyperbolic differential operators*, Trans. Amer. Math. Soc. **333** (1992), 821–833.

[133] P. Stefanov and G. Vodev, *Distribution of resonances for the Neumann problem in linear elasticity outside a strictly convex body*, Duke Math. J. **78** (1995), 677–714.

[134] L. Tartar, *H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A **115** (1990), 193–230.

[135] D. Tataru, *Unique continuation for solutions to PDE's; between Hörmander's theorem and Holmgren's theorem*, Comm. Partial Differential Equations **20** (1995), 855–884.

[136] _____, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl. **75** (1996), 367–408.

[137] _____, *Unique continuation for operators with partially analytic coefficients*, J. Math. Pures Appl. **78** (1999), 505–521.

[138] M. Tucsnak and G. Weiss, *Observation and control for operator semigroups*, Birkhäuser Advanced Texts: Basel Textbooks (2009).

[139] J. Vancostenoble and E. Zuazua, *Null controllability for the heat equation with singular inverse-square potentials*, J. Funct. Anal. **254** (2008), 1864–1902.

[140] W. Wang, *Carleman inequalities and unique continuation for higher-order elliptic differential operators*, Duke Math. J. **74** (1994), 107–128.

[141] T. H. Wolff, *A property of measures in $\mathbb{R}^N$ and an application to unique continuation*, Geom. Funct. Anal. **2** (1992), 225–284.

[142] X. Zhang, *A unified controllability/observability theory for some stochastic and deterministic partial differential equations*, Proceedings of the International Congress of Mathematicians Volume IV (2010), 3008–3034.

[143] C. Zuily, *Uniqueness and nonuniqueness in the Cauchy problem*, Progress in Mathematics **33** (1983) Birkhäuser Boston Inc.

Laboratoire de Mathématiques de Versailles, Université de Versailles St Quentin, CNRS UMR 8100, 45, Avenue des États-Unis, 78035 Versailles, France.

E-mail: luc.robbiano@uvsq.fr

Models and feedback stabilization of open quantum systems

Pierre Rouchon

Abstract. At the quantum level, feedback-loops have to take into account measurement back-action. We present here the structure of the Markovian models including such back-action and sketch two stabilization methods: measurement-based feedback where an open quantum system is stabilized by a classical controller; coherent or autonomous feedback where a quantum system is stabilized by a quantum controller with decoherence (reservoir engineering). We begin to explain these models and methods for the photon box experiments realized in the group of Serge Haroche (Nobel Prize 2012). We present then these models and methods for general open quantum systems.

Mathematics Subject Classification (2010). Primary 93B52, 93D15, 81V10, 81P15; Secondary 93C20, 81P68, 35Q84.

Keywords. Markov model, open quantum system, quantum filtering, quantum feedback, quantum master equation.

1. Introduction

Serge Haroche has obtained the Physics Nobel Prize in 2012 for a series of crucial experiments on observations and manipulations of photons with atoms. The book [33], written with Jean-Michel Raimond, describes the physics (Cavity Quantum Electro-Dynamics, CQED) underlying these experiments done at Laboratoire Kastler Brossel (LKB). These experimental setups, illustrated on figure 2.1 and named in the sequel "the LKB photon box", rely on fundamental examples of open quantum systems constructed with harmonic oscillators and qubits. Their time evolutions are captured by stochastic dynamical models based on three features, specific to the quantum world and listed below.

1. The state of a quantum system is described either by the wave function $|\psi\rangle$ a vector of length one belonging to some separable Hilbert space $\mathcal{H}$ of finite or infinite dimension, or, more generally, by the density operator ρ that is a non-negative Hermitian operator on $\mathcal{H}$ with trace one. When the system can be described by a wave function $|\psi\rangle$ (pure state), the density operator ρ coincides with the orthogonal projector on the line spanned by $|\psi\rangle$ and $\rho = |\psi\rangle\langle\psi|$ with usual Dirac notations. In general the rank of ρ exceeds one, the state is then mixed and cannot be described by a wave function. When the system is closed, the time evolution of $|\psi\rangle$ is governed by the Schrödinger equation
$$\frac{d}{dt}|\psi\rangle = -\tfrac{i}{\hbar}\boldsymbol{H}|\psi\rangle \tag{1.1}$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

where $\boldsymbol{H}$ is the system Hamiltonian, an Hermitian operator on $\mathcal{H}$ that could possibly depend on time t via some time-varying parameters (classical control inputs). When the system is closed, the evolution of ρ is governed by the Liouville/von-Neumann equation

$$\frac{d}{dt}\rho = -\tfrac{i}{\hbar}\big[\boldsymbol{H},\rho\big] = -\tfrac{i}{\hbar}\big(\boldsymbol{H}\rho - \rho\boldsymbol{H}\big). \tag{1.2}$$

2. Dissipation and irreversibility has its origin in the "collapse of the wave packet" induced by the measurement. A measurement on the quantum system of state $|\psi\rangle$ or ρ is associated of an observable $\boldsymbol{O}$, an Hermitian operator on $\mathcal{H}$, with spectral decomposition $\sum_\mu \lambda_\mu \boldsymbol{P}_\mu$: $\boldsymbol{P}_\mu$ is the orthogonal projector on the eigen-space associated to the eigen-value λ_μ. The measurement process attached to $\boldsymbol{O}$ is assumed to be instantaneous and obeys to the following rules:

   - the measurement outcome μ is obtained with probability $p_\mu = \langle\psi|\boldsymbol{P}_\mu|\psi\rangle$ or $p_\mu = \mathrm{Tr}\,(\rho\boldsymbol{P}_\mu)$, depending on the state $|\psi\rangle$ or ρ just before the measurement;
   - just after the measurement process, the quantum state is changed to $|\psi\rangle_+$ or ρ_+ according to the mappings

$$|\psi\rangle \mapsto |\psi\rangle_+ = \frac{\boldsymbol{P}_\mu|\psi\rangle}{\sqrt{\langle\psi|\boldsymbol{P}_\mu|\psi\rangle}} \quad \text{or} \quad \rho \mapsto \rho_+ = \frac{\boldsymbol{P}_\mu\rho\boldsymbol{P}_\mu}{\mathrm{Tr}\,(\rho\boldsymbol{P}_\mu)}$$

   where μ is the observed measurement outcome. These mappings describe the measurement back-action and have no classical counterpart.

3. Most systems are composite systems built with several sub-systems. The quantum states of such composite systems live in the tensor product of the Hilbert spaces of each sub-system. This is a crucial difference with classical composite systems where the state space is built with Cartesian products. Such tensor products have important implications such as entanglement with existence of non separable states. Consider a bi-partite system made of two sub-systems: the sub-system of interest S with Hilbert space $\mathcal{H}_S$ and the measured sub-system M with Hilbert space $\mathcal{H}_M$. The quantum state of this bi-partite system (S,M) lives in $\mathcal{H} = \mathcal{H}_S \otimes \mathcal{H}_M$. Its Hamiltonian $\boldsymbol{H}$ is constructed with the Hamiltonians of the sub-systems, $\boldsymbol{H}_S$ and $\boldsymbol{H}_M$, and an interaction Hamiltonian $\boldsymbol{H}_{int}$ made of a sum of tensor products of operators (not necessarily Hermitian) on S and M:

$$\boldsymbol{H} = \boldsymbol{H}_S \otimes \boldsymbol{I}_M + \boldsymbol{H}_{int} + \boldsymbol{I}_S \otimes \boldsymbol{H}_M$$

   with $\boldsymbol{I}_S$ and $\boldsymbol{I}_M$ identity operators on $\mathcal{H}_S$ and $\mathcal{H}_M$, respectively. The measurement operator $\boldsymbol{O} = \boldsymbol{I}_S \otimes \boldsymbol{O}_M$ is here a simple tensor product of identity on S and the Hermitian operator $\boldsymbol{O}_M$ on $\mathcal{H}_M$, since only M is directly measured. Its spectrum is degenerate: the multiplicities of the eigenvalues are necessarily greater or equal to the dimension of $\mathcal{H}_S$.

This paper shows that, despite different mathematical formulations, dynamical models describing open quantum systems admit the same structure, essentially given by the Markov model (3.2), and directly derived from the three quantum features listed here above. Section 2 explains the construction of such Markov models for the LKB photon box and its stabilization by measurement-based and coherent feedbacks. These stabilizing feedbacks

rely on control Lyapunov functions, quantum filtering and reservoir engineering. The next sections explain these models and methods for general open quantum systems. In section 3 (resp. section 4) general discrete-time (resp. continuous-time) systems are considered. In appendix, operators, key states and formulae are presented for the quantum harmonic oscillator and for the qubit, two key quantum systems. These notations are used and not explicitly recalled throughout sections 2, 3 and 4.

2. The LKB photon box

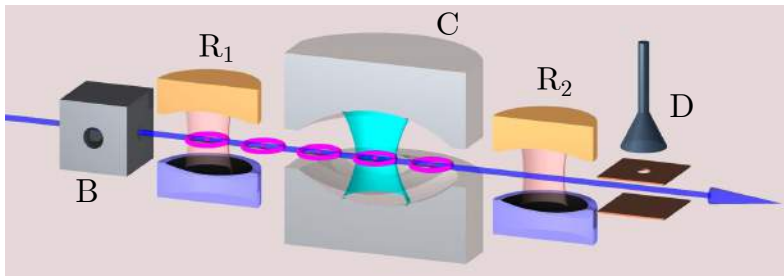


Figure 2.1. Scheme of the LBK experiment where photons are observed via probe atoms. The photons in blue are trapped between the two mirrors of the cavity C. They are probed by two-level atoms (the small pink torus) flying out the preparation box B, passing through the cavity C and measured in D. Each atom is manipulated before and after C in Ramsey cavities R_1 and R_2, respectively. It is finally detected in D either in ground state $|g\rangle$ or in excited state $|e\rangle$.

2.1. The ideal Markov model. The LKB photon box of figure 2.1, a bi-partite system with the photons as first sub-system and the probe atom as second sub-system, illustrates in an almost perfect and fundamental way the three quantum features listed in the introduction section. This system is a discrete time system with sampling period τ around $80\ \mu s$, the time interval between probe atoms. Step $k \in \mathbb{N}$ corresponds to time $t = k\tau$. At $t = k\tau$, the photons are assumed to be described by the wave function $|\psi\rangle_k$ of an harmonic oscillator (see appendix A). At $t = k\tau$, the probe atom number k, modeled as a qubit (see appendix B), gets outside the box B in ground state $|g\rangle$. Between $t \in [k\tau, (k+1)\tau[$, the wave function $|\Psi\rangle$ of this composite system, photons/atom number k, is governed by a Schrödinger evolution

$$\frac{d}{dt}|\Psi\rangle = -\frac{i}{\hbar}\boldsymbol{H}|\Psi\rangle$$

with starting condition $|\Psi\rangle_{k\tau} = |\psi\rangle_k \otimes |g\rangle$ and where $\boldsymbol{H}$ is the photons/atom Hamiltonian depending possibly on t. Appendix C presents typical Hamiltonians in the resonant and dispersive cases. We have thus a propagator between $t = k\tau$ and $t = (k+1)\tau^-$, $U_{(k\tau,(k+1)\tau^-)}$, from which we get $|\Psi\rangle$ at time $t = (k+1)\tau^-$, just before detector D where the energy of the atom is measured via $\boldsymbol{O} = \boldsymbol{I}_S \otimes \boldsymbol{\sigma_z}$. The following relation,

$$|\Psi\rangle_{(k+1)\tau^-} = U_{(k\tau,(k+1)\tau^-)}|\psi\rangle_k \otimes |g\rangle \triangleq \boldsymbol{M}_g|\psi\rangle_k \otimes |g\rangle + \boldsymbol{M}_e|\psi\rangle_k \otimes |e\rangle,$$

valid for any $|\psi\rangle_k$, defines the measurement operators $\boldsymbol{M}_g$ and $\boldsymbol{M}_e$ on the Hilbert space of the photons $\mathcal{H}_S$. Since, for all $|\psi\rangle_k$, $|\Psi\rangle_{(k+1)\tau^-}$ is of length 1, we have necessarily

$\boldsymbol{M}_g^\dagger \boldsymbol{M}_g + \boldsymbol{M}_e^\dagger \boldsymbol{M}_e = \boldsymbol{I}_S$. At time $t = (k+1)\tau^-$, we measure $\boldsymbol{O} = \lambda_e \boldsymbol{I}_S \otimes |e\rangle\langle e| + \lambda_g \boldsymbol{I}_S \otimes |g\rangle\langle g|$ with two highly degenerate eigenvalues $\lambda_e = 1$, $\lambda_g = -1$ of eigenspaces $\mathcal{H}_S \otimes |e\rangle$ and $\mathcal{H}_S \otimes |g\rangle$, respectively. According to the measurement quantum rules, we can get only two outcomes μ, either $\mu = g$ or $\mu = e$. With outcome μ, just after the measurement, at time $(k+1)\tau$ the quantum state $|\Psi\rangle$ is changed to

$$|\Psi\rangle_{(k+1)\tau^-} \mapsto |\Psi\rangle_{(k+1)\tau} = \frac{\boldsymbol{M}_\mu |\psi\rangle_k}{\sqrt{\langle\psi_k|\boldsymbol{M}_\mu^\dagger \boldsymbol{M}_\mu|\psi_k\rangle}} \otimes |\mu\rangle.$$

Moreover the probability to get μ is $\langle\psi_k|\boldsymbol{M}_\mu^\dagger \boldsymbol{M}_\mu|\psi_k\rangle$. Since $|\Psi\rangle_{(k+1)\tau}$ is now a simple tensor product (separate state), we can forget the atom number k and summarize the evolution of the photon wave function between $t = k\tau$ and $t = (k+1)\tau$ by the following Markov process

$$|\psi\rangle_{k+1} = \begin{cases} \frac{\boldsymbol{M}_g|\psi\rangle_k}{\sqrt{\langle\psi_k|\boldsymbol{M}_g^\dagger \boldsymbol{M}_g|\psi_k\rangle}}, & \text{with probability } \langle\psi_k|\boldsymbol{M}_g^\dagger \boldsymbol{M}_g|\psi_k\rangle; \\ \frac{\boldsymbol{M}_e|\psi\rangle_k}{\sqrt{\langle\psi_k|\boldsymbol{M}_e^\dagger \boldsymbol{M}_g|\psi_k\rangle}}, & \text{with probability } \langle\psi_k|\boldsymbol{M}_e^\dagger \boldsymbol{M}_e|\psi_k\rangle. \end{cases}$$

More generally, for an arbitrary quantum state ρ_k of the photons at step k, we have

$$\rho_{k+1} = \begin{cases} \frac{\boldsymbol{M}_g \rho_k \boldsymbol{M}_g^\dagger}{\mathrm{Tr}\left(\boldsymbol{M}_g \rho_k \boldsymbol{M}_g^\dagger\right)}, & \text{with probability } p_g(\rho_k) = \mathrm{Tr}\left(\boldsymbol{M}_g \rho_k \boldsymbol{M}_g^\dagger\right); \\ \frac{\boldsymbol{M}_e \rho_k \boldsymbol{M}_e^\dagger}{\mathrm{Tr}\left(\boldsymbol{M}_e \rho_k \boldsymbol{M}_e^\dagger\right)}, & \text{with probability } p_e(\rho_k) = \mathrm{Tr}\left(\boldsymbol{M}_e \rho_k \boldsymbol{M}_e^\dagger\right). \end{cases} \tag{2.1}$$

The measurement operators $\boldsymbol{M}_g$ and $\boldsymbol{M}_e$ are implicitly defined by the Schrödinger propagator between $k\tau$ and $(k+1)\tau$. They always satisfy $\boldsymbol{M}_g^\dagger \boldsymbol{M}_g + \boldsymbol{M}_e^\dagger \boldsymbol{M}_e = \boldsymbol{I}_S$.

2.2. Quantum non demolition (QND) measurement. For a well tuned composite evolution $U_{(k\tau,(k+1)\tau^-)}$ (see [33]) with a dispersive interaction, one get the following measurement operators, functions of the photon-number operator $\boldsymbol{N}$,

$$\boldsymbol{M}_g = \cos\left(\tfrac{\phi_0 \boldsymbol{N} + \phi_R}{2}\right), \quad \boldsymbol{M}_e = \sin\left(\tfrac{\phi_0 \boldsymbol{N} + \phi_R}{2}\right) \tag{2.2}$$

where ϕ_0 and ϕ_R are tunable real parameters. The Markov process (2.1) admits then a lot of interesting properties characterizing QND measurement.

- For any function $g : \mathbb{R} \mapsto \mathbb{R}$, $V_g(\rho) = \mathrm{Tr}\left(g(\boldsymbol{N})\rho\right)$ is a martingale:

  $$\mathbb{E}\left(V_g(\rho_{k+1}) \,/\, \rho_k\right) = V_g(\rho_k)$$

  where $\mathbb{E}\left(x \,/\, y\right)$ stands for conditional expectation of x knowing y. This results from elementary properties of the trace and from the commutation of $\boldsymbol{M}_g$ and $\boldsymbol{M}_e$ with $\boldsymbol{N}$.
- For any integer $\bar{n}$, the photon-number state $|\bar{n}\rangle\langle\bar{n}|$ ($\bar{n} \in \mathbb{N}$) is a steady-state: any realization of (2.1) starting from $\rho_0 = |\bar{n}\rangle\langle\bar{n}|$ is constant: $\forall k \geq 0$, $\rho_k \equiv |\bar{n}\rangle\langle\bar{n}|$.
- When (ϕ_R, ϕ_0, π) are $\mathbb{Q}$-independent, there is no other steady state than these photon-number states. Moreover, for any initial density operator ρ_0 with a finite photon-number support ($\rho_0|m\rangle = 0$ for m large enough), the probability that ρ_k converges towards the steady state $|\bar{n}\rangle\langle\bar{n}|$ is $\mathrm{Tr}\left(|\bar{n}\rangle\langle\bar{n}|\rho_0\right) = \langle\bar{n}|\rho_0|\bar{n}\rangle$. Since $\mathrm{Tr}\left(\rho_0\right) = 1 = \sum_{\bar{n}\in\mathbb{N}} \langle\bar{n}|\rho_0|\bar{n}\rangle$, the Markov process (2.1) converges almost surely towards a photon-number state, whatever its initial state ρ_0 is.

The proof of this convergence result is essentially based on a Lyapunov function, a supermartingale, $V(\rho) = -\sum_{n\in\mathbb{N}}\langle n\,|\rho|\,n\rangle^2$. Simple computations yield

$$\mathbb{E}\left(V(\rho_{k+1})\,/\,\rho_k\right) = V(\rho_k) - Q(\rho_k)$$

where $Q(\rho) \geq 0$ is given by the following formula

$$Q(\rho) = \sum_{n\in\mathbb{N}} \frac{\sin^2(\phi_0 n+\phi_R)}{4} \left(\frac{\cos^2\left(\frac{\phi_0 n+\phi_R}{2}\right)\langle n|\rho|n\rangle}{\sum_{n'}\cos^2\left(\frac{\phi_0 n'+\phi_R}{2}\right)\langle n'|\rho|n'\rangle} - \frac{\sin^2\left(\frac{\phi_0 n+\phi_R}{2}\right)\langle n|\rho|n\rangle}{\sum_{n'}\sin^2\left(\frac{\phi_0 n'+\phi_R}{2}\right)\langle n'|\rho|n'\rangle} \right)^2$$

Since (ϕ_0, ϕ_R, π) are $\mathbb{Q}$-independent, $Q(\rho) = 0$ implies that, for some $\bar{n} \in \mathbb{N}$, $\rho = |\bar{n}\rangle\langle\bar{n}|$. One concludes then with usual probability and compactness arguments [39], despite the fact that the underlying Hilbert space is of infinite dimension. Other and also more precise results can be found in [9].

2.3. Stabilization of photon-number states by feedback. Take $\bar{n} \in \mathbb{N}$. With measurement operators (2.2), the Markov process (2.1) admits $\bar{\rho} = |\bar{n}\rangle\langle\bar{n}|$ as steady state. We describe here the measurement-based feedback (quantum-state feedback) implemented experimentally in [56] and that stabilizes $\bar{\rho}$. Here the scalar classical control input u consists in applying, just after the atom measurement in D, a coherent displacement of tunable amplitude u. This yields the following control Markov process

$$\rho_{k+1} = \begin{cases} \dfrac{\boldsymbol{D}_{u_k}\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger\boldsymbol{D}_{u_k}^\dagger}{\mathrm{Tr}\left(\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger\right)} & y_k = g \text{ with probability } p_{g,k} = \mathrm{Tr}\left(\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger\right) \\ \dfrac{\boldsymbol{D}_{u_k}\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\boldsymbol{D}_{u_k}^\dagger}{\mathrm{Tr}\left(\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right)} & y_k = e \text{ with probability } p_{e,k} = \mathrm{Tr}\left(\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right) \end{cases} \tag{2.3}$$

where $u_k \in \mathbb{R}$ is the control at step k, $\boldsymbol{D}_u = e^{u\boldsymbol{a}^\dagger - u\boldsymbol{a}}$ is the displacement of amplitude u(see appendix A) and y_k is the measurement outcome at step k.

The stabilization of $\bar{\rho}$ is based on a state-feedback function f, $u = f(\rho)$, such that almost all closed-loop trajectories of (2.3) with $u_k = f(\rho_k)$ converge towards $\bar{\rho}$ for any initial condition ρ_0. The construction of f exploits the open-loop martingales $\mathrm{Tr}\,(g(\boldsymbol{N})\rho)$ to construct the following strict control Lyapunov function:

$$V_\epsilon(\rho) = \sum_n \left(-\epsilon\langle n\,|\rho|\,n\rangle^2 + \sigma_n\langle n|\rho|n\rangle\right)$$

where $\epsilon > 0$ is small enough and

$$\sigma_n = \begin{cases} \frac{1}{4} + \sum_{\nu=1}^{\bar{n}} \frac{1}{\nu} - \frac{1}{\nu^2}, & \text{if } n = 0; \\ \sum_{\nu=n+1}^{\bar{n}} \frac{1}{\nu} - \frac{1}{\nu^2}, & \text{if } n \in [1, \bar{n}-1]; \\ 0, & \text{if } n = \bar{n}; \\ \sum_{\nu=\bar{n}+1}^{n} \frac{1}{\nu} + \frac{1}{\nu^2}, & \text{if } n \in [\bar{n}+1, +\infty]. \end{cases}$$

The weight σ_n are all non negative, $n \mapsto \sigma_n$ is strictly decreasing (resp. increasing) for $n \leq \bar{n}$ (resp. $n \geq \bar{n}$) and minimum for $n = \bar{n}$. The feedback law $u = f(\rho)$ is obtained by

choosing u such that the expectation value of $V_\epsilon(\rho_{k+1})$, knowing $\rho_k = \rho$ and $u_k = u$, is as small as possible:

$$u = f(\rho) =: \underset{v\in[-\bar{u},\bar{u}]}{\text{Argmin}} \quad V_\epsilon\Big(\boldsymbol{D}_v\left(\boldsymbol{M}_g\rho\boldsymbol{M}_g^\dagger + \boldsymbol{M}_e\rho\boldsymbol{M}_e^\dagger\right)\boldsymbol{D}_v^\dagger\Big)$$

where $\bar{u} > 0$ is some prescribed bound on $|u|$. Such a feedback law achieves global stabilization since, in closed-loop, the Lyapunov function is strict:

$$\forall \rho \neq |\bar{n}\rangle\langle\bar{n}|, \quad V_\epsilon\Big(\boldsymbol{D}_{f(\rho)}\left(\boldsymbol{M}_g\rho\boldsymbol{M}_g^\dagger + \boldsymbol{M}_e\rho\boldsymbol{M}_e^\dagger\right)\boldsymbol{D}_{f(\rho)}^\dagger\Big) < V_\epsilon\Big(\rho\Big).$$

Formal convergence proofs can be found in [3] for any finite dimensional approximations resulting from a truncation to a finite number of photons and in [59] for the infinite dimension.

2.4. A more realistic Markov model with detection errors. The experimental implementation of the above feedback law [56] has to cope with several sources of imperfections. We focus here on measurement errors and show how the Markov process has to be changed to take into account these errors. Assume that we know the detection error rates characterized by $\mathbb{P}(y = e/\mu = g) = \eta_g \in [0,1]$ (resp. $\mathbb{P}(y = g/\mu = e) = \eta_e \in [0,1]$) the probability of erroneous assignation to e (resp. g) when the atom collapses in g (resp. e). Without error, the quantum state ρ_k obeys to (2.1). A direct application of Bayes law provides the expectation of ρ_{k+1}, knowing ρ_k and the effective detector signal y_k, possibly corrupted by a detection error. When $y_k = g$, this expectation value is given by $\frac{(1-\eta_g)\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger+\eta_e\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger}{\mathrm{Tr}\left((1-\eta_g)\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger+\eta_e\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right)}$ and, when $y_k = e$, by $\frac{\eta_g\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger+(1-\eta_e)\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger}{\mathrm{Tr}\left(\eta_g\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger+(1-\eta_e)\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right)}$. Moreover the probability to get $y_k = g$ is $\mathrm{Tr}\left((1-\eta_g)\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger + \eta_e\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right)$ and to get $y_k = e$ is $\mathrm{Tr}\left(\eta_g\boldsymbol{M}_g\rho\boldsymbol{M}_g^\dagger + (1-\eta_e)\boldsymbol{M}_e\rho\boldsymbol{M}_e^\dagger\right)$. This means that the Markov process (2.1) must be changed to

$$\rho_{k+1} = \begin{cases} \frac{(1-\eta_g)\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger+\eta_e\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger}{\mathrm{Tr}\left((1-\eta_g)\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger+\eta_e\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right)} & \text{when } y_k = g, \\ \frac{\eta_g\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger+(1-\eta_e)\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger}{\mathrm{Tr}\left(\eta_g\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger+(1-\eta_e)\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right)} & \text{when } y_k = e, \end{cases} \tag{2.4}$$

with $\mathrm{Tr}\left((1-\eta_g)\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger + \eta_e\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right)$ and $\mathrm{Tr}\left(\eta_g\boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger + (1-\eta_e)\boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger\right)$ being the probabilities to detect $y_k = g$ and e, respectively. The quantum state ρ_k is thus a conditional state: it is the expectation value of the projector associated to the photon wave function at step k, knowing its value at step $k = 0$ and the detection outcomes $(y_0, \ldots, y_{k-1})$.

All other experimental imperfections including decoherence can be treated in the same way (see, e.g., [26, 58]) and yield to a quantum state governed by a Markov process with a similar structure. In fact all usual models of open quantum systems admit the same structure, either in discrete-time (see section 3) or in continuous-time (see section 4).

2.5. The real-time stabilization algorithm. Let us give more details on the real-time implementation used in [56] of this quantum-state feedback. The sampling period τ is around $80\ \mu s$. The controller set-point is an integer $\bar{n}$ labelling the steady-state $\bar{\rho} = |\bar{n}\rangle\langle\bar{n}|$ to be stabilized. At time step k, the real-time computer

1. reads y_k the measurement outcome for probe atom k;

2. updates the quantum state from previous step value ρ_{k-1} to ρ_k using y_k and a Markov model slightly more complicated but of same structure as (2.4); this update corresponds to a quantum filter (see subsection 3.3).
3. computes u_k as $f(\rho_k)$ (state feedback) where f results from minimizing the expectation of the control Lyapunov function $V_\epsilon(\rho)$ at step $k+1$, knowing ρ_k;
4. send via an antenna a micro-wave pulse calibrated to obtain the displacement $\boldsymbol{D}_{u_k}$ on the photons.

All the details of this quantum feedback are given in [55]. In particular, the Markov model takes into account several experimental imperfections such as finite life-time of the photons (around $1/10\ s$) and a delay of 5 steps in the feedback loop. Convergence results related to this feedback scheme are given in [3].

2.6. Reservoir engineering stabilization of Schrödinger cats. It is possible to stabilize the photons trapped in cavity C (figure 2.1) without any such measurement-based feedback, just by well tuned interactions with the probe atoms and without measuring them in D. Such kind of stabilization, known as reservoir engineering [50], can be seen as a generalization of optical pumping techniques [37]. Such stabilization methods are illustrative of coherent (or autonomous) feedback where the controller is an open quantum system. In [53], a realistic implementation of such passive stabilization method is proposed. It stabilizes a coherent superposition of classical photon-states with opposite phases, a Schrödinger phase-cats with wave functions of the form $(|\alpha\rangle + i|\text{-}\alpha\rangle)/\sqrt{2}$, where $|\alpha\rangle$ is the coherent state of amplitude $\alpha \in \mathbb{R}$. We explain here the convergence analysis of such passive stabilization using the notations and operator definitions given in appendix A.

The atom entering the cavity C is prepared through R_1 in a partially excited state $\cos(u/2)|g\rangle + \sin(u/2)|e\rangle$ with $u \in [0, \pi/2[$ (south hemisphere of the Bloch sphere). Its interaction with the photons is first dispersive with positive detuning during its entrance, then resonant in the cavity middle and finally dispersive with negative detuning when leaving the cavity. The resulting measurement operators $\boldsymbol{M}_g$ and $\boldsymbol{M}_e$ appearing in (2.1) admit then the following form (see [54] for detailed derivations):

$$\boldsymbol{M}_g = e^{-i\widetilde{h}(\boldsymbol{N})}\widetilde{\boldsymbol{M}}_g e^{i\widetilde{h}(\boldsymbol{N})}, \quad \boldsymbol{M}_e = e^{-i\widetilde{h}(\boldsymbol{N})}\widetilde{\boldsymbol{M}}_e e^{i\widetilde{h}(\boldsymbol{N})}$$

with $n \mapsto \widetilde{h}(n)$ a real function, with $\boldsymbol{I}$ standing for $\boldsymbol{I}_S$, with

$$\widetilde{\boldsymbol{M}}_g = \cos(\tfrac{u}{2})\cos\left(\tfrac{\theta(\boldsymbol{N})}{2}\right) + \epsilon\sin(\tfrac{u}{2})\frac{\sin\left(\frac{\theta(\boldsymbol{N})}{2}\right)}{\sqrt{\boldsymbol{N}}}\,\boldsymbol{a}^\dagger$$

$$\widetilde{\boldsymbol{M}}_e = \sin(\tfrac{u}{2})\cos\left(\tfrac{\theta(\boldsymbol{N}+\boldsymbol{I})}{2}\right) - \epsilon\cos(\tfrac{u}{2})\,\boldsymbol{a}\frac{\sin\left(\frac{\theta(\boldsymbol{N})}{2}\right)}{\sqrt{\boldsymbol{N}}}$$

and with $n \mapsto \theta(n)$ a real function such that

$$\theta(0) = 0,\ \forall n > 0,\ \theta(n) \in]0, \pi[\text{ and } \lim_{n\mapsto+\infty}\theta(n) = \pi/2$$

Since we do not measure the atoms, the photon state ρ_{k+1} at step $k+1$ is given by the following recurrence from the state ρ_k at step k:

$$\rho_{k+1} = \boldsymbol{K}(\rho_k) \triangleq \boldsymbol{M}_g\rho_k\boldsymbol{M}_g^\dagger + \boldsymbol{M}_e\rho_k\boldsymbol{M}_e^\dagger.$$

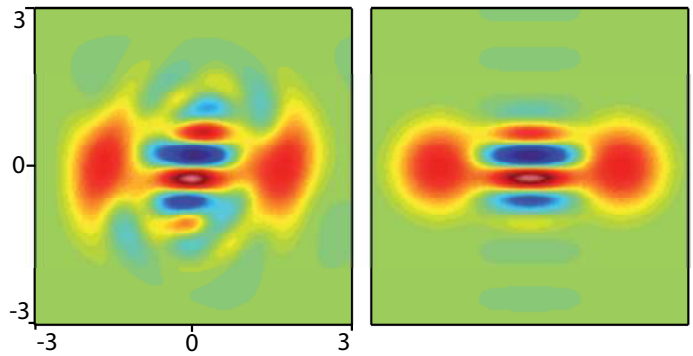


Figure 2.2. Left: Wigner function of ρ_∞ stabilized by reservoir engineering in [54]. Right: Wigner function of a prefect Schrödinger phase-cat, $\frac{1}{2}\Big(|\alpha_\infty\rangle + i|\text{-}\alpha_\infty\rangle\Big)\Big(\langle\alpha_\infty| + i\langle\text{-}\alpha_\infty|\Big)$, with an average number of photons identical to ρ_∞ ($\alpha_\infty = \sqrt{\mathrm{Tr}\,(\boldsymbol{N}\rho_\infty)}$). The color map is identical to figure A.1.

Consider the change of frame associated to the unitary transformation $e^{-i\widetilde{h}(\boldsymbol{N})}$: $\rho = e^{-i\widetilde{h}(\boldsymbol{N})} \times \widetilde{\rho} e^{i\widetilde{h}(\boldsymbol{N})}$. Then we have $\widetilde{\rho}_{k+1} = \widetilde{\boldsymbol{K}}(\widetilde{\rho}_k) \triangleq \widetilde{\boldsymbol{M}}_g \widetilde{\rho}_k (\widetilde{\boldsymbol{M}}_g)^\dagger + \widetilde{\boldsymbol{M}}_e \widetilde{\rho}_k (\widetilde{\boldsymbol{M}}_e)^\dagger$. It is proved in [40] that, since $|u| \leq \pi/2$, exists a unique common eigen-state $|\widetilde{\psi}\rangle \in \mathcal{H}_S$ of $\widetilde{\boldsymbol{M}}_g$ and $\widetilde{\boldsymbol{M}}_e$. Thus $\widetilde{\rho}_\infty = |\widetilde{\psi}\rangle\langle\widetilde{\psi}|$ is a fixed point of $\widetilde{\boldsymbol{K}}$. It is also proved in [40] that the $\widetilde{\rho}_k$'s converge to $\widetilde{\rho}_\infty$ when the function θ is strictly increasing. Since the underlying Hilbert space $\mathcal{H}_S$ is of infinite dimension, it is important to precise the type of convergence. For any initial condition $\widetilde{\rho}_0$ such that $\mathrm{Tr}\,(\boldsymbol{N}\widetilde{\rho}_0) < +\infty$, then $\lim_{k\mapsto+\infty} \mathrm{Tr}\,\big((\widetilde{\rho}_k - \widetilde{\rho}_\infty)^2\big) = 0$ (Frobenius norm on Hilbert-Schmidt operators). Since $\mathrm{Tr}\,(\boldsymbol{N}\rho) \equiv \mathrm{Tr}\,(\boldsymbol{N}\widetilde{\rho})$, we have the convergence of ρ_k towards $\rho_\infty = e^{-i\widetilde{h}(\boldsymbol{N})}\widetilde{\rho}_\infty e^{i\widetilde{h}(\boldsymbol{N})}$ as soon as the initial energy $\mathrm{Tr}\,(\boldsymbol{N}\rho_0)$ is finite: $\lim_{k\mapsto+\infty} \mathrm{Tr}\,\Big(\big((\rho_k - \rho_\infty)^2\Big) = 0$. When θ is not strictly increasing, we conjecture that such convergence towards ρ_∞ still holds true.

For well chosen experimental parameters [54], $\widetilde{\rho}_\infty$ is close to a coherent state $|\alpha_\infty\rangle\langle\alpha_\infty|$ for some $\alpha_\infty \in \mathbb{R}$ and $\widetilde{h}(\boldsymbol{N}) \approx \pi\boldsymbol{N}^2/2$. Since

$$e^{-i\frac{\pi}{2}\boldsymbol{N}^2}|\alpha_\infty\rangle = \frac{e^{-i\pi/4}}{\sqrt{2}}\big(|\alpha_\infty\rangle + i|\text{-}\alpha_\infty\rangle\big),$$

we have under realistic conditions $\lim_{k\mapsto+\infty}\rho_k \approx \frac{1}{2}\Big(|\alpha_\infty\rangle + i|\text{-}\alpha_\infty\rangle\Big)\Big(\langle\alpha_\infty| - i\langle\text{-}\alpha_\infty|\Big)$, a coherent superposition of the classical states $|\alpha_\infty\rangle$ and $|\text{-}\alpha_\infty\rangle$ of same amplitude but of opposite phases, i.e. a Schrödinger phase-cat. Figure 2.2 displays numerical computations of the Wigner function of ρ_∞ obtained with realistic parameters.

3. Discrete-time systems

The theory of open quantum systems starts with the contributions of Davies [25]. The goal of this section is first to present in an elementary way the general structure of the Markov models describing such systems. Some related stabilization problems are also addressed. Throughout this section, $\mathcal{H}$ is an Hilbert space; for each time-step $k \in \mathbb{N}$, ρ_k denotes the

density operator describing the state of the quantum Markov process; for all k, ρ_k is an Hilbert-Schmidt operator on $\mathcal{H}$, Hermitian and of trace one; the set of continuous operators on $\mathcal{H}$ is denoted by $\mathcal{L}(\mathcal{H})$; expectation values are denoted by the symbol $\mathbb{E}(\ )$.

3.1. Markov models. Take a positive integer m and consider a finite set $(\boldsymbol{M}_\mu)_{\mu\in\{1,\dots,m\}}$ of operators on $\mathcal{H}$ such that

$$\boldsymbol{I} = \sum_{\mu=1}^{m} \boldsymbol{M}_\mu^\dagger \boldsymbol{M}_\mu \tag{3.1}$$

where $\boldsymbol{I}$ is the identity operator. Then each $\boldsymbol{M}_\mu \in \mathcal{L}(\mathcal{H})$. Take another positive integer m' and consider a left stochastic $m' \times m$-matrix $(\eta_{\mu'\mu})$: its entries are non-negative and $\forall \mu \in \{1,\dots,m\}$, $\sum_{\mu'=1}^{m'} \eta_{\mu'\mu} = 1$. Consider the Markov process of state ρ and output $y \in \{1,\dots,m'\}$ (measurement outcome) defined via the transition rule

$$\rho_{k+1} = \frac{\sum_\mu \eta_{\mu'\mu} \boldsymbol{M}_\mu \rho_k \boldsymbol{M}_\mu^\dagger}{\mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu} \boldsymbol{M}_\mu \rho_k \boldsymbol{M}_\mu^\dagger\right)}, \quad y_k = \mu' \text{ with probability } \mathbb{P}_{\mu'}(\rho_k) \tag{3.2}$$

where $\mathbb{P}_{\mu'}(\rho) = \mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu} \boldsymbol{M}_\mu \rho \boldsymbol{M}_\mu^\dagger\right)$. At each step k, the probability to have $y_k = \mu'$ depends on the quantum state ρ_k and is given by $\mathbb{P}_{\mu'}(\rho_k)$.

3.2. Kraus and unital maps. The Kraus map $\boldsymbol{K}$ corresponds to the master equation of (3.2). It is given by the expectation value of ρ_{k+1} knowing ρ_k:

$$\boldsymbol{K}(\rho) \triangleq \sum_\mu \boldsymbol{M}_\mu \rho \boldsymbol{M}_\mu^\dagger = \mathbb{E}\left(\rho_{k+1} \,/\, \rho_k = \rho\right). \tag{3.3}$$

In quantum information [47] such Kraus maps describe quantum channels. They admit many interesting properties. In particular, they are contractions for many metrics (see [49] for the characterization, in finite dimension, of metrics for which any Kraus map is a contraction). We just recall below two such metrics. For any density operators ρ and σ we have

$$D(\boldsymbol{K}(\rho), \boldsymbol{K}(\sigma)) \le D(\rho,\sigma) \text{ and } F(\boldsymbol{K}(\rho), \boldsymbol{K}(\sigma)) \ge F(\rho,\sigma) \tag{3.4}$$

where the trace distance D and fidelity F are given by

$$D(\rho,\sigma) \triangleq \mathrm{Tr}\left(|\rho-\sigma|\right) \text{ and } F(\rho,\sigma) \triangleq \mathrm{Tr}^2\left(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right). \tag{3.5}$$

Fidelity is between 0 and 1: $F(\rho,\sigma) = 1$ if and only if, $\rho = \sigma$. Moreover $F(\rho,\sigma) = F(\sigma,\rho)$. If $\sigma = |\psi\rangle\langle\psi|$ is a pure state ($|\psi\rangle$ element of $\mathcal{H}$ of length one), $F(\rho,\sigma)$ coincides with the Frobenius product: $F(\rho,|\psi\rangle\langle\psi|) \equiv \mathrm{Tr}\left(\rho|\psi\rangle\langle\psi|\right) = \langle\psi|\rho|\psi\rangle$. Kraus maps provide the evolution of open quantum systems from an initial state ρ_0 without information coming from the measurements (see [33, chapter 4: the environment is watching]):

$$\rho_{k+1} = \boldsymbol{K}(\rho_k) \text{ for } k = 0, 1, \dots, .$$

This corresponds to the "Schrödinger description" of the dynamics.

The "Heisenberg description" is given by the dual map $\boldsymbol{K}^*$. It is characterized by

$$\mathrm{Tr}\left(A\boldsymbol{K}(\rho)\right) = \mathrm{Tr}\left(\boldsymbol{K}^*(A)\rho\right)$$

and defined for any operator A on $\mathcal{H}$ by

$$\boldsymbol{K}^*(A) = \sum_\mu \boldsymbol{M}_\mu^\dagger A \boldsymbol{M}_\mu.$$

Technical conditions on A are required when $\mathcal{H}$ is of infinite dimension, they are not given here (see, e.g., [25]). The map $\boldsymbol{K}^*$ is unital since (3.1) reads $\boldsymbol{K}^*(\boldsymbol{I}) = \boldsymbol{I}$. As $\boldsymbol{K}$, the dual map $\boldsymbol{K}^*$ admits a lot of interesting properties. It is noticed in [57] that, based on a theorem due of Birkhoff [14], such unital maps are contractions on the cone of non-negative Hermitian operators equipped with the Hilbert's projective metric. In particular, when $\mathcal{H}$ is of finite dimension, we have, for any Hermitian operator A:

$$\lambda_{min}(A) \leq \lambda_{min}(\boldsymbol{K}^*(A)) \leq \lambda_{max}(\boldsymbol{K}^*(A)) \leq \lambda_{max}(A)$$

where λ_{min} and λ_{max} correspond to the smallest and largest eigenvalues. As shown in [51], such contraction properties based on Hilbert's projective metric have important implications in quantum information theory.

To emphasize the difference between the "Schrödinger description" and the "Heisenberg description" of the dynamics, let us translate convergence issues from the "Schrödinger description" to the "Heisenberg one". Assume, for clarity's sake, that $\mathcal{H}$ is of finite dimension. Suppose also that $\boldsymbol{K}$ admits the density operator $\bar\rho$ as unique fixed point and that, for any initial density operator ρ_0, the density operator at step k, ρ_k, defined by k iterations of $\boldsymbol{K}$, converges towards $\bar\rho$ when k tends to ∞. Then $k \mapsto D(\rho_k, \bar\rho)$ is decreasing and converges to 0 whereas $k \mapsto F(\rho_k, \bar\rho)$ is increasing and converges to 1.

The translation of this convergence in the "Heisenberg description" is the following: for any initial operator A_0, its k iterates via $\boldsymbol{K}^*$, A_k, converge towards $\mathrm{Tr}\left(A_0\bar\rho\right)\boldsymbol{I}$. Moreover when A_0 is Hermitian, $k \mapsto \lambda_{min}(A_k)$ and $k \mapsto \lambda_{max}(A_k)$ are respectively increasing and decreasing and both converge to $\mathrm{Tr}\left(A_0\bar\rho\right)$.

3.3. Quantum filtering. Quantum filtering has its origin in Belavkin's work [13] on continuous-time open quantum systems (see section 4). The state ρ_k of (3.2) is not directly measured: open quantum systems are governed by hidden-state Markov model. Quantum filtering provides an estimate ρ_k^{est} of ρ_k based on an initial guess ρ_0^{est} (possibly different from ρ_0) and the measurement outcomes y_l between 0 and $k-1$:

$$\rho_{l+1}^{\text{est}} = \frac{\sum_\mu \eta_{y_l\mu} \boldsymbol{M}_\mu \rho_l^{\text{est}} \boldsymbol{M}_\mu^\dagger}{\mathrm{Tr}\left(\sum_\mu \eta_{y_l\mu} \boldsymbol{M}_\mu \rho_l^{\text{est}} \boldsymbol{M}_\mu^\dagger\right)}, \quad l \in \{0, \ldots, k-1\}. \tag{3.6}$$

Thus $(\rho, \rho^{\text{est}})$ is the state of an extended Markov process governed by the following rule

$$\rho_{k+1} = \frac{\sum_\mu \eta_{\mu'\mu} \boldsymbol{M}_\mu \rho_k \boldsymbol{M}_\mu^\dagger}{\mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu} \boldsymbol{M}_\mu \rho_k \boldsymbol{M}_\mu^\dagger\right)} \text{ and } \rho_{k+1}^{\text{est}} = \frac{\sum_\mu \eta_{\mu'\mu} \boldsymbol{M}_\mu \rho_k^{\text{est}} \boldsymbol{M}_\mu^\dagger}{\mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu} \boldsymbol{M}_\mu \rho_k^{\text{est}} \boldsymbol{M}_\mu^\dagger\right)}$$

with transition probability $\mathbb{P}_{\mu'}(\rho_k) = \mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu} \boldsymbol{M}_\mu \rho_k \boldsymbol{M}_\mu^\dagger\right)$ depending on ρ_k and independent of ρ_k^{est}.

When $\mathcal{H}$ is of finite dimension, it is shown in [58] with an inequality proved in [52] that such discrete-time quantum filters are always stable in the following sense: the fidelity between ρ and its estimate ρ^{est} is a sub-martingale for any initial condition ρ_0 and ρ_0^{est};

$$\mathbb{E}\left(F(\rho_{k+1},\rho_{k+1}^{\text{est}}) \mid (\rho_k,\rho_k^{\text{est}})\right) \geq F(\rho_k,\rho_k^{\text{est}})$$

This result does not guaranty that ρ_k^{est} converges to ρ_k when k tends to infinity. The convergence characterization of ρ^{est} towards ρ via checkable conditions on the left stochastic matrix $(\eta_{\mu'\mu})$ and on the set of operators $(\boldsymbol{M}_\mu)$ remains an open problem [60, 61].

3.4. Stabilization via measurement-based feedback. Assume now that the operators $\boldsymbol{M}_\mu$ appearing in (3.2) and satisfying (3.1), depend also on a control input u belonging to some admissible set $\mathcal{U}$ (typically a discrete set or a compact subset of $\mathbb{R}^p$ for some positive integer p). Then we have the following control Markov model with input $u \in \mathcal{U}$, hidden state ρ and measured output $y \in \{1,\ldots,m'\}$:

$$\rho_{k+1} = \frac{\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(u_k)\rho_k\boldsymbol{M}_\mu^\dagger(u_k)}{\mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(u_k)\rho_k\boldsymbol{M}_\mu^\dagger(u_k)\right)},\ y_k=\mu' \text{ with probability } \mathbb{P}_{\mu'}(\rho_k,u_k) \tag{3.7}$$

where $\mathbb{P}_{\mu'}(\rho,u) = \mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(u)\rho\boldsymbol{M}_\mu^\dagger(u)\right)$. Assume that for some nominal admissible input $\bar{u} \in \mathcal{U}$, this Markov process admits a steady state $\bar{\rho}$. This means that, for any $\mu' \in \{1,\ldots,m'\}$ we have $\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(\bar{u})\bar{\rho}\boldsymbol{M}_\mu^\dagger(\bar{u}) = \mathbb{P}_{\mu'}(\bar{\rho},\bar{u})\bar{\rho}$. The measurement-based feedback stabilization of the steady-state $\bar{\rho}$ is the following problem: for any initial condition ρ_0, find for any $k \in \mathbb{N}$ a control input $u_k \in \mathcal{U}$ depending only on ρ_0 and on the past y values, $(y_0,\ldots,y_{k-1})$, such that ρ_k converges almost surely towards $\bar{\rho}$.

Quantum-state feedback scheme, $u = f(\rho)$, can be used here. They can be based on Lyapunov techniques. Potential candidates of Lyapunov functions $V(\rho)$ could be related to the metrics for which the open-loop Kaus map with $\bar{u}$ is contracting. Specific V depending on the precise structure of the system could be more adapted as for the LKB photon box [3]. Such Lyapunov feedback laws are then given by the minimization versus $u \in \mathcal{U}$ of $\mathbb{E}\left(V(\rho_{k+1}) \mid \rho_k=\rho, u_k=u\right)$.

Assume that we have a stabilizing feedback law $u = f(\rho)$: $\bar{u} = f(\bar{\rho})$ and the trajectories of (3.7) with $u_k = f(\rho_k)$ converge almost surely towards $\bar{\rho}$. Since ρ is not directly accessible, one has to replace ρ_k by its estimate ρ_k^{est} to obtain u_k. Experimental implementations of such quantum feedback laws admit necessarily an observer/controller structure governed by a Markov process of state (ρ,ρ^{est}) with the following transition rule:

$$\begin{aligned}
\rho_{k+1} &= \frac{\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(f(\rho_k^{\text{est}}))\rho_k\boldsymbol{M}_\mu^\dagger(f(\rho_k^{\text{est}}))}{\mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(f(\rho_k^{\text{est}}))\rho_k\boldsymbol{M}_\mu^\dagger(f(\rho_k^{\text{est}}))\right)}\\
\rho_{k+1}^{\text{est}} &= \frac{\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(f(\rho_k^{\text{est}}))\rho_k^{\text{est}}\boldsymbol{M}_\mu^\dagger(f(\rho_k^{\text{est}}))}{\mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(f(\rho_k^{\text{est}}))\rho_k^{\text{est}}\boldsymbol{M}_\mu^\dagger(f(\rho_k^{\text{est}}))\right)}
\end{aligned} \tag{3.8}$$

with probability $\mathbb{P}_{\mu'}(\rho_k,f(\rho_k^{\text{est}})) = \mathrm{Tr}\left(\sum_\mu \eta_{\mu'\mu}\boldsymbol{M}_\mu(f(\rho_k^{\text{est}}))\rho_k\boldsymbol{M}_\mu^\dagger(f(\rho_k^{\text{est}}))\right)$ depending on ρ_k and ρ_k^{est}. In [16] a separation principle is proved with elementary arguments (see also [3]): if $\mathcal{H}$ is of finite dimension, if $\bar{\rho}$ is a pure state ($\bar{\rho} = |\bar{\psi}\rangle\langle\bar{\psi}|$ for some $|\bar{\psi}\rangle$ in $\mathcal{H}$)

and if $\mathrm{Ker}(\rho_0^{\mathrm{est}}) \subset \mathrm{Ker}(\rho_0)$, then almost all realizations of (3.8) converge to the steady-state $(\bar\rho, \bar\rho)$. The stabilizing feedback schemes used in experiments [56] and [64] exploit such observer/controller structure and rely on this separation principle where the designs of the stabilizing feedback (controller) and of the quantum-state filter (observer) are done separately.

With such feedback scheme we loose the linear formulation of the ensemble-average master equation with a Kraus map. In general, there is no simple formulation of the master equation governing the expectation value of ρ_k in closed-loop. Nevertheless, for systems where the measurement step producing the output y_k is followed by a control action characterized by u_k, it is possible via a static output feedback, $u_k = f(y_k)$ where f is now some function from $\{1, \ldots, m'\}$ to $\mathcal{U}$, to preserve in closed-loop such Kraus-map formulations. These specific feedback schemes, called Markovian feedbacks, are due to Wiseman and have important applications. They are well explained and illustrated in the recent book [63].

3.5. Stabilization of pure states by reservoir engineering. With T as sampling period, a possible formalization of this passive stabilization method is as follows. The goal is to stabilize a pure state $\bar\rho_S = |\bar\psi_S\rangle\langle\bar\psi_S|$ for a system S with Hilbert space $\mathcal{H}_S$ and Hamiltonian operator H_S ($|\bar\psi_S\rangle \in \mathcal{H}_S$ is of length one). To achieve this goal consider a "realistic" quantum controller of Hilbert space $\mathcal{H}_C$ with initial state $|\theta_C\rangle$ and with Hamiltonian $\boldsymbol{H}_C$. One has to design an adapted interaction between S and C with a well chosen interaction Hamiltonian $\boldsymbol{H}_{int}$, an Hermitian operator on $\mathcal{H}_{S,C} = \mathcal{H}_S \otimes \mathcal{H}_C$. This controller C and its interaction with S during the sampling interval of length T have to fulfill the conditions explained below in order to stabilize $\bar\rho_S$.

Denote by $\boldsymbol{U}_{S,C} = \boldsymbol{U}(T)$ the propagator between 0 and time T for the composite system (S, C): $\boldsymbol{U}(t)$ is the unitary operator on $\mathcal{H}_{S,C}$ defined by

$$\frac{d}{dt}\boldsymbol{U} = -\tfrac{i}{\hbar}\Big(\boldsymbol{H}_S \otimes \boldsymbol{I}_C + \boldsymbol{H}_{int} + \boldsymbol{I}_S \otimes \boldsymbol{H}_C\Big)\boldsymbol{U}, \quad \boldsymbol{U}(0) = \boldsymbol{I}_{S,C}$$

where $\boldsymbol{I}_S$, $\boldsymbol{I}_C$ and $\boldsymbol{I}_{S,C}$ are the identity operators on $\mathcal{H}_S$, $\mathcal{H}_C$, and $\mathcal{H}_{S,C}$, respectively. To the propagator $U_{S,C}$ and the initial controller wave function $|\theta_C\rangle \in \mathcal{H}_C$ is attached a Kraus map $\boldsymbol{K}$ on $\mathcal{H}_S$,

$$\boldsymbol{K}(\rho_S) = \sum_\mu \boldsymbol{M}_\mu \rho_S \boldsymbol{M}_\mu^\dagger$$

where the operators $\boldsymbol{M}_\mu$ on $\mathcal{H}_S$ are defined by the decomposition,

$$\forall |\psi_S\rangle \in \mathcal{H}_S, \quad \boldsymbol{U}_{S,C}\big(|\psi_S\rangle \otimes |\theta_C\rangle\big) = \sum_\mu \big(\boldsymbol{M}_\mu |\psi_S\rangle\big) \otimes |\lambda_\mu\rangle,$$

with $(|\lambda_\mu\rangle)$ any ortho-normal basis of $\mathcal{H}_C$. Despite the fact that the operators $(\boldsymbol{M}_\mu)$ depend on the choice of this basis, the Kraus map $\boldsymbol{K}$ is independent of this choice: it depends only on $U_{S,C}$ and $|\theta_C\rangle$.

The first stabilization condition is the following: the Kraus operators $\boldsymbol{M}_\mu$ have to admit $|\bar\psi_S\rangle$ as common a eigen-vector since $\bar\rho_S$ has to be a fixed point of $\boldsymbol{K}$ ($\boldsymbol{K}(\bar\rho_S) = \bar\rho_S$).

The second stabilization condition is the following: for any initial density operator $\rho_{S,0}$, the iterates $\rho_{S,k}$ of $\boldsymbol{K}$ converge to $\bar\rho_S$, i.e.,

$$\lim_{k \mapsto +\infty} \rho_{S,k} = \bar\rho_S \text{ where } \rho_{S,k} = \boldsymbol{K}(\rho_{S,k-1}).$$

When these two conditions are satisfied, the repetition of the same interaction for each sampling interval $[kT,(k+1)T]$ $(k \in \mathbb{N})$ with a controller-state $|\theta_C\rangle$ at kT ensures that the density operator of S at kT, $\rho_{S,k}$, converges to $\bar{\rho}_S$ since $\rho_{S,k} = \boldsymbol{K}(\rho_{S,k-1})$. Here, the so-called reservoir is made of the infinite set of identical controller systems C indexed by $k \in \mathbb{N}$, with initial state $|\theta_C\rangle$ and interacting sequentially with S during $[kT,(k+1)T]$.

4. Continuous-time systems

4.1. Stochastic master equations. These models have their origins in the work of Davies [25], are related to quantum trajectories [18, 24] and are connected to Belavkin quantum filters [13]. A modern and mathematical exposure of the diffusive models is given in [5]. These models are interpreted here as continuous-time versions of (3.2). They are based on stochastic differential equations, also called Stochastic Master Equations (SME). They provide the evolution of the density operator ρ_t with respect to the time t. They are driven by a finite number of independent Wiener processes indexed by ν, $(W_{\nu,t})$, each of them being associated to a continuous classical and real signal, $y_{\nu,t}$, produced by detector ν. These SMEs admit the following form:

$$d\rho_t = \left(-\tfrac{i}{\hbar}[\boldsymbol{H},\rho_t] + \sum_\nu \boldsymbol{L}_\nu \rho_t \boldsymbol{L}_\nu^\dagger - \tfrac{1}{2}(\boldsymbol{L}_\nu^\dagger \boldsymbol{L}_\nu \rho_t + \rho_t \boldsymbol{L}_\nu^\dagger \boldsymbol{L}_\nu) \right) dt + \sum_\nu \sqrt{\eta_\nu} \left(\boldsymbol{L}_\nu \rho_t + \rho_t \boldsymbol{L}_\nu^\dagger - \operatorname{Tr}\left((\boldsymbol{L}_\nu + \boldsymbol{L}_\nu^\dagger)\rho_t \right) \rho_t \right) dW_{\nu,t} \quad (4.1)$$

where $\boldsymbol{H}$ is the Hamiltonian operator on the underlying Hilbert space $\mathcal{H}$ and $\boldsymbol{L}_\nu$ are arbitrary operators (not necessarily Hermitian) on $\mathcal{H}$. Each measured signal $y_{\nu,t}$ is related to ρ_t and $W_{\nu,t}$ by the following output relationship:

$$dy_{\nu,t} = dW_{\nu,t} + \sqrt{\eta_\nu} \operatorname{Tr}\left((\boldsymbol{L}_\nu + \boldsymbol{L}_\nu^\dagger)\,\rho_t \right) dt$$

where $\eta_\nu \in [0,1]$ is the efficiency of detector ν. The ensemble average of ρ_t obeys thus to a linear differential equation, also called master or Lindblad-Kossakowski differential equation [38, 41]:

$$\frac{d}{dt}\rho = -\tfrac{i}{\hbar}[\boldsymbol{H},\rho] + \sum_\nu \boldsymbol{L}_\nu \rho_t \boldsymbol{L}_\nu^\dagger - \tfrac{1}{2}(\boldsymbol{L}_\nu^\dagger \boldsymbol{L}_\nu \rho_t + \rho_t \boldsymbol{L}_\nu^\dagger \boldsymbol{L}_\nu). \quad (4.2)$$

It is the continuous-time analogue of the Kraus map $\boldsymbol{K}$ associated to the Markov process (2.4).

In fact (3.2) and (4.1) have the same structure. This becomes obvious if one remarks that, with standard Itō rules, (4.1) admits the following formulation

$$\rho_{t+dt} = \frac{\boldsymbol{M}_{dy_t} \rho_t \boldsymbol{M}_{dy_t}^\dagger + \sum_\nu (1-\eta_\nu) \boldsymbol{L}_\nu \rho_t \boldsymbol{L}_\nu^\dagger dt}{\operatorname{Tr}\left(\boldsymbol{M}_{dy_t} \rho_t \boldsymbol{M}_{dy_t}^\dagger + \sum_\nu (1-\eta_\nu) \boldsymbol{L}_\nu \rho_t \boldsymbol{L}_\nu^\dagger dt \right)}$$

with $\boldsymbol{M}_{dy_t} = \boldsymbol{I} + \left(-\tfrac{i}{\hbar}\boldsymbol{H} - \tfrac{1}{2}\sum_\nu \boldsymbol{L}_\nu^\dagger \boldsymbol{L}_\nu \right) dt + \sum_\nu \sqrt{\eta_\nu} dy_{\nu t} \boldsymbol{L}_\nu$. With such a formulation, it becomes clear that (4.1) preserves the trace and the non-negativeness of ρ. This formulation

provides also directly a time discretization numerical scheme preserving non-negativeness of ρ

Mixed diffusive/jump stochastic master equations can be considered. Additional Poisson counting processes $(N_\mu(t))$ are added in parallel to the Wiener processes $(W_{\nu,t})$ [2]:

$$
\begin{aligned}
d\rho_t = &\left(-\tfrac{i}{\hbar}[\boldsymbol{H},\rho_t] + \sum_\nu \boldsymbol{L}_\nu \rho_t \boldsymbol{L}_\nu^\dagger - \tfrac{1}{2}(\boldsymbol{L}_\nu^\dagger \boldsymbol{L}_\nu \rho_t + \rho_t \boldsymbol{L}_\nu^\dagger \boldsymbol{L}_\nu) \right) dt \\
&+ \sum_\nu \sqrt{\eta_\nu} \Big(\boldsymbol{L}_\nu \rho_t + \rho_t \boldsymbol{L}_\nu^\dagger - \operatorname{Tr}\left((\boldsymbol{L}_\nu + \boldsymbol{L}_\nu^\dagger)\rho_t \right) \rho_t \Big) dW_{\nu,t} \\
&+ \left(\sum_\mu \boldsymbol{V}_\mu \rho_t \boldsymbol{V}_\mu^\dagger - \tfrac{1}{2}(\boldsymbol{V}_\mu^\dagger \boldsymbol{V}_\mu \rho_t + \rho_t \boldsymbol{V}_\mu^\dagger \boldsymbol{V}_\mu) \right) dt \\
+ \sum_\mu \left(\frac{\overline{\theta}_\mu \rho_t + \sum_{\mu'} \overline{\eta}_{\mu,\mu'} \boldsymbol{V}_\mu \rho_t \boldsymbol{V}_\mu^\dagger}{\overline{\theta}_\mu + \sum_{\mu'} \overline{\eta}_{\mu,\mu'} \operatorname{Tr}\left(\boldsymbol{V}_{\mu'} \rho_t \boldsymbol{V}_{\mu'}^\dagger \right)} - \rho_t \right) &\left(dN_\mu(t) - \left(\overline{\theta}_\mu + \sum_{\mu'} \overline{\eta}_{\mu,\mu'} \operatorname{Tr}\left(\boldsymbol{V}_{\mu'} \rho_t \boldsymbol{V}_{\mu'}^\dagger \right) \right) dt \right)
\end{aligned}
$$

where the $\boldsymbol{V}_\mu$'s are operators on $\mathcal{H}$, where the additional parameters $\overline{\theta}_\mu, \overline{\eta}_{\mu,\mu'} \geq 0$ with $\overline{\eta}_{\mu'} = \sum_\mu \overline{\eta}_{\mu,\mu'} \leq 1$, describe counting imperfections. For each μ, $\left(\overline{\theta}_\mu + \sum_{\mu'} \overline{\eta}_{\mu,\mu'} \operatorname{Tr}\left(\boldsymbol{V}_{\mu'} \rho_t \boldsymbol{V}_{\mu'}^\dagger \right) \right) dt$ is the probability to increment by one N_μ between t and $t+dt$. The above stochastic model is similar to the discrete-time Markov process (3.2). The transition from ρ_t to ρ_{t+dt} is given by the following two possibilities:

- either, for some μ, $dN_\mu(t) = N_\mu(t+dt) - N_\mu(t) = 1$, then we have the transition $\rho_{t+dt} = \frac{\overline{\theta}_\mu \rho_t + \sum_{\mu'} \overline{\eta}_{\mu,\mu'} \boldsymbol{V}_{\mu'} \rho_t \boldsymbol{V}_{\mu'}^\dagger}{\overline{\theta}_\mu + \sum_{\mu'} \overline{\eta}_{\mu,\mu'} \operatorname{Tr}\left(\boldsymbol{V}_{\mu'} \rho_t \boldsymbol{V}_{\mu'}^\dagger \right)}$;
- or, $\forall \mu,\ dN_\mu(t) = 0$, and we have the transition

$$
\rho_{t+dt} = \frac{\boldsymbol{M}_{dy_t} \rho_t \boldsymbol{M}_{dy_t}^\dagger + \sum_\nu (1-\eta_\nu) \boldsymbol{L}_\nu \rho_t \boldsymbol{L}_\nu^\dagger dt + \sum_\mu (1-\overline{\eta}_\mu) \boldsymbol{V}_\mu \rho_t \boldsymbol{V}_\mu^\dagger dt}{\operatorname{Tr}\left(\boldsymbol{M}_{dy_t} \rho_t \boldsymbol{M}_{dy_t}^\dagger + \sum_\nu (1-\eta_\nu) \boldsymbol{L}_\nu \rho_t \boldsymbol{L}_\nu^\dagger dt + \sum_\mu (1-\overline{\eta}_\mu) \boldsymbol{V}_\mu \rho_t \boldsymbol{V}_\mu^\dagger dt \right)}
$$

with

$$
\begin{aligned}
\boldsymbol{M}_{dy_t} = \boldsymbol{I} + &\left(-\tfrac{i}{\hbar}\boldsymbol{H} - \tfrac{1}{2}\sum_\nu \boldsymbol{L}_\nu^\dagger \boldsymbol{L}_\nu + \tfrac{1}{2}\sum_\mu \left(\overline{\eta}_\mu \operatorname{Tr}\left(\boldsymbol{V}_\mu \rho_t \boldsymbol{V}_\mu^\dagger \right) \boldsymbol{I} - \boldsymbol{V}_\mu^\dagger \boldsymbol{V}_\mu \right) \right) dt \\
&+ \sum_\nu \sqrt{\eta_\nu} dy_{\nu t} \boldsymbol{L}_\nu
\end{aligned}
$$

and $dy_{\nu,t} = \sqrt{\eta_\nu} \operatorname{Tr}\left((\boldsymbol{L}_\nu + \boldsymbol{L}_\nu^\dagger)\rho_t \right) dt + dW_{\nu,t}$.

Such transition relationships can be exploited by numerical integration schemes in order to preserve positiveness of ρ. In particular, when all η_ν, $\overline{\theta}_\mu$ and $\overline{\eta}_{\mu,\mu'}$ are equal to zero, we recover, up to second order terms, the explicit Euler numerical scheme for the Lindblad-Kossakovski equation.

4.2. Quantum filtering. For clarity's sake, take in (4.1) a single measurement y_t associated to operator $\boldsymbol{L}$, detection efficiency $\eta \in [0,1]$ and scalar Wiener process W_t: $dy_t = \sqrt{\eta}\,\mathrm{Tr}\left((\boldsymbol{L}+\boldsymbol{L}^\dagger)\,\rho_t\right)\,dt + dW_t$. The continuous-time counterpart of (3.6) provides the estimate ρ_t^{est} by the Belavkin quantum filtering process

$$\begin{aligned} d\rho_t^{\mathrm{est}} &= -\tfrac{i}{\hbar}[\boldsymbol{H},\rho_t^{\mathrm{est}}]\,dt + \left(\boldsymbol{L}\rho_t^{\mathrm{est}}\boldsymbol{L}^\dagger - \tfrac{1}{2}(\boldsymbol{L}^\dagger\boldsymbol{L}\rho_t^{\mathrm{est}} + \rho_t^{\mathrm{est}}\boldsymbol{L}^\dagger\boldsymbol{L})\right)\,dt \\ &\quad + \sqrt{\eta}\left(\boldsymbol{L}\rho_t^{\mathrm{est}} + \rho_t^{\mathrm{est}}\boldsymbol{L}^\dagger - \mathrm{Tr}\left((\boldsymbol{L}+\boldsymbol{L}^\dagger)\rho_t^{\mathrm{est}}\right)\rho_t^e\right)\left(dy_t - \sqrt{\eta}\,\mathrm{Tr}\left((\boldsymbol{L}+\boldsymbol{L}^\dagger)\rho_t^{\mathrm{est}}\right)\,dt\right). \end{aligned}$$

initialized to any density matrix ρ_0^{est}. Thus $(\rho,\rho^{\mathrm{est}})$ obeys to the following set of nonlinear stochastic differential equations

$$\begin{aligned} d\rho_t &= -\tfrac{i}{\hbar}[\boldsymbol{H},\rho_t]\,dt + \left(\boldsymbol{L}\rho_t\boldsymbol{L}^\dagger - \tfrac{1}{2}(\boldsymbol{L}^\dagger\boldsymbol{L}\rho_t + \rho_t\boldsymbol{L}^\dagger\boldsymbol{L})\right)\,dt \\ &\quad + \sqrt{\eta}\left(\boldsymbol{L}\rho_t + \rho_t\boldsymbol{L}^\dagger - \mathrm{Tr}\left((\boldsymbol{L}+\boldsymbol{L}^\dagger)\rho_t\right)\rho_t\right)\,dW_t \end{aligned}$$

$$\begin{aligned} d\rho_t^{\mathrm{est}} &= -\tfrac{i}{\hbar}[\boldsymbol{H},\rho_t^{\mathrm{est}}]\,dt + \left(\boldsymbol{L}\rho_t^{\mathrm{est}}\boldsymbol{L}^\dagger - \tfrac{1}{2}(\boldsymbol{L}^\dagger\boldsymbol{L}\rho_t^{\mathrm{est}} + \rho_t^{\mathrm{est}}\boldsymbol{L}^\dagger\boldsymbol{L})\right)\,dt \\ &\quad + \sqrt{\eta}\left(\boldsymbol{L}\rho_t^{\mathrm{est}} + \rho_t^{\mathrm{est}}\boldsymbol{L}^\dagger - \mathrm{Tr}\left((\boldsymbol{L}+\boldsymbol{L}^\dagger)\rho_t^{\mathrm{est}}\right)\rho_t^{\mathrm{est}}\right)\,dW_t \\ &\quad + \eta\left(\boldsymbol{L}\rho_t^{\mathrm{est}} + \rho_t^{\mathrm{est}}\boldsymbol{L}^\dagger - \mathrm{Tr}\left((\boldsymbol{L}+\boldsymbol{L}^\dagger)\rho_t^{\mathrm{est}}\right)\rho_t^{\mathrm{est}}\right)\mathrm{Tr}\left((\boldsymbol{L}+\boldsymbol{L}^\dagger)(\rho_t-\rho_t^{\mathrm{est}})\right)\,dt. \end{aligned}$$

It is proved in [2] that such filtering process is always stable in the sense that, as for the discrete-time case, the fidelity between ρ_t and ρ_t^{est} is a sub-martingale. In [61] a first convergence analysis of these filters is proposed. Nevertheless the convergence characterization in terms of the operators $\boldsymbol{H}$, $\boldsymbol{L}$ and the parameter η remains an open problem as far as we know.

Formulations of quantum filters for stochastic master equations driven by an arbitrary number of Wiener and Poisson processes can be found in [2].

4.3. Stabilization via measurement-based feedback. Assume that the Hamiltonian $H = H_0 + uH_1$ appearing in (4.2) depends on some scalar control input u, H_0 and H_1 being Hermitian operators on $\mathcal{H}$. Assume also that $\bar\rho = |\bar\psi\rangle\langle\bar\psi|$ is a steady-state of (4.2) for $u = 0$. Necessarily $|\bar\psi\rangle$ is an eigen-vectors of each $\boldsymbol{L}_\nu$, $\boldsymbol{L}_\nu|\bar\psi\rangle = \lambda_\nu|\bar\psi\rangle$ for some $\lambda_\nu \in \mathbb{C}$. This implies that $\bar\rho$ is also a steady-state of (4.1) with $u = 0$, since $\boldsymbol{L}_\nu\bar\rho + \bar\rho\boldsymbol{L}_\nu^\dagger = \mathrm{Tr}\left((\boldsymbol{L}_\nu + \boldsymbol{L}_\nu^\dagger)\bar\rho\right)\bar\rho$. The stabilization of $\bar\rho$ consists then in finding a feedback law $u = f(\rho)$ with $f(\bar\rho) = 0$ such that almost all trajectories ρ_t of the closed-loop system (4.1) with $H = H(t) = H_0 + f(\rho_t)H_1$ converge to $\bar\rho$ when t tends to $+\infty$. Such feedback law could be obtained by Lyapunov techniques as in [46]. As in the discrete-case, ρ_t is replaced, in the feedback law, by its estimate ρ_t^{est} obtained via quantum filtering. Convergence is then guarantied as soon as $\mathrm{Ker}\,\rho_0^{\mathrm{est}} \subset \mathrm{Ker}\,\rho_0$ [16]. Other feedback schemes not relying directly on the quantum state ρ_t but still based on past values of the measurement signals y_ν can be considered (see [63] for Markovian feedbacks; see [17, 62] for recent experimental implementations).

4.4. Stabilization via coherent feedback. This passive stabilization method has its origin, for classical system, in the classical Watt regulator where a mechanical system, the steam

machine, was controlled by another mechanical system, a conical pendulum. As initially shown in [44], the study of such closed-loop systems highlights stability and convergence as the main mathematical issues. For quantum systems, these issues remain similar and are related to reservoir engineering [42, 50].

As in the discrete-time case, the goal remains to stabilize a pure state $\bar{\rho}_S = |\bar{\psi}_S\rangle\langle\bar{\psi}_S|$ for system S (Hilbert space $\mathcal{H}_S$ and Hamiltonian $\boldsymbol{H}_S$) by coupling to the controller system C (Hilbert space $\mathcal{H}_C$, Hamiltonian $\boldsymbol{H}_C$) via the interaction $\boldsymbol{H}_{int}$, an Hermitian operator on $\mathcal{H}_S \otimes \mathcal{H}_C$. The controller C is subject to decoherence described by the set $\left(\boldsymbol{L}_{C,\nu}\right)$ of operators on $\mathcal{H}_C$ indexed by ν. The closed-loop system is a composite system with Hilbert space $\mathcal{H} = \mathcal{H}_S \otimes \mathcal{H}_C$. Its density operator ρ obeys to (4.2) with $\boldsymbol{H} = \boldsymbol{H}_S \otimes \boldsymbol{I}_C + \boldsymbol{I}_S \otimes \boldsymbol{H}_C + \boldsymbol{H}_{int}$ and $\boldsymbol{L}_\nu = \boldsymbol{I}_S \otimes \boldsymbol{L}_{C,\nu}$ ($\boldsymbol{I}_S$ and $\boldsymbol{I}_C$ identity operators on $\mathcal{H}_S$ and $\mathcal{H}_C$, respectively). Stabilization is achieved when $\rho(t)$ converges, whatever its initial condition $\rho(0)$ is, to a separable state of the form $\bar{\rho}_S \otimes \bar{\rho}_C$ where $\bar{\rho}_C$ could possibly depend on t and/or on $\rho(0)$. In several interesting cases, such as cooling [32], coherent feedback is shown to outperform measurement-based feedback.

The asymptotic analysis (stability and convergence rates) for such composite closed-loop systems is far from being obvious, even if such analysis is based on known properties for each subsystem and for the coupling Hamiltonian $\boldsymbol{H}_{int}$. When $\mathcal{H}$ is of infinite dimension, convergence analysis becomes more difficult. To have an idea of the mathematical issues, take the harmonic oscillators considered in [45]. They are nonlinearly coupled to coherent drives. These open quantum systems could have important applications for quantum computations and are governed by the following kind of master equations:

$$\frac{d}{dt}\rho = u[(\boldsymbol{a}^\dagger)^r - \boldsymbol{a}^r, \rho] + \kappa\left((\boldsymbol{a}^r\rho(\boldsymbol{a}^\dagger)^r - \tfrac{1}{2}(\boldsymbol{a}^\dagger)^r\boldsymbol{a}^r\rho - \tfrac{1}{2}\rho(\boldsymbol{a}^\dagger)^r\boldsymbol{a}^r\right) \tag{4.3}$$

where $u > 0$ and $\kappa > 0$ are constant parameters and r is an integer greater than 1. Set $\bar{\alpha} = \sqrt[r]{2u/\kappa}$ and for $s \in \{0, 1, \ldots, r-1\}$, $\bar{\alpha}_s = e^{2is\pi/r}\bar{\alpha}$. Denote by $|\bar{\alpha}_s\rangle$ the coherent state of complex amplitude $\bar{\alpha}_s$. Computations exploiting properties of coherent states recalled in appendix A show that, for any s, $|\bar{\alpha}_s\rangle\langle\bar{\alpha}_s|$ is a steady state of (4.3). Moreover the set of steady states corresponds to the density operators $\bar{\rho}$ with support inside the vector space spanned by the $|\bar{\alpha}_s\rangle$ for $s \in \{0, 1, \ldots, r-1\}$. We conjecture that, for initial conditions $\rho(0)$ with finite energy ($\mathrm{Tr}\,(\rho\boldsymbol{N}) < \infty$), the solutions of (4.3) are well defined and converge in Frobenius norm to such steady states $\bar{\rho}$ possibly depending on $\rho(0)$. Having sharp estimations of the convergence rates is also an open question. We cannot apply here the existing general convergence results towards "full rank steady-states" (see, e.g., [4][chapter 4]): here the rank of such steady states $\bar{\rho}$ is at most r. Another formulation of such dynamics can be given via the Wigner function W^ρ of ρ (see appendix A). With the correspondence (A.2), (4.3) yields a partial differential equation describing the time evolution of W^ρ: this equation is of order one in time but of order $2r$ versus the two variables of the phase plane. It corresponds to a Fokker-Planck equation of high order.

5. Concluding remarks

The above exposure deals with specific and limited aspects of modelling and control of open quantum systems. It does not consider many other interesting developments such as

- controllability and motion planing in finite dimension [23, 31] and in infinite dimension (see, e.g., [10–12, 21, 27]);
- quantum Langevin equations and input/output approach [28], quantum signal amplification [22] and linear quantum systems [35];
- (S, L, H) formalism for quantum networks [30];
- master equations and quantum Fokker Planck equations [19, 20];
- optimal control methods [7, 8, 15, 29, 48].

More topics can also be found in the review articles [1, 34, 43].

A. Quantum harmonic oscillator

We just recall here some useful formulae (see, e.g., [6]). The Hamiltonian formulation of the classical harmonic oscillator of pulsation $\omega > 0$, $\frac{d^2}{dt^2}x = -\omega^2 x$, is as follows:

$$\frac{d}{dt}x = \omega p = \frac{\partial H}{\partial p}, \quad \frac{d}{dt}p = -\omega x = -\frac{\partial H}{\partial x}$$

with the classical Hamiltonian $H(x,p) = \frac{\omega}{2}(p^2 + x^2)$. The correspondence principle yields the following quantization: H becomes an operator $\boldsymbol{H}$ on the function of $x \in \mathbb{R}$ with complex values. The classical state $(x(t), p(t))$ is replaced by the quantum state $|\psi\rangle_t$ associated to the function $\psi(x,t) \in \mathbb{C}$. At each t, $\mathbb{R} \ni x \mapsto \psi(x,t)$ is measurable and $\int_{\mathbb{R}} |\psi(x,t)|^2 dx = 1$: for each t, $|\psi\rangle_t \in L^2(\mathbb{R}, \mathbb{C})$.

The Hamiltonian $\boldsymbol{H}$ is derived from the classical one H by replacing x by the Hermitian operator $\boldsymbol{X} \equiv \frac{x}{\sqrt{2}}$ and p by the Hermitian operator $\boldsymbol{P} \equiv -\frac{\imath}{\sqrt{2}}\frac{\partial}{\partial x}$:

$$\frac{\boldsymbol{H}}{\hbar} = \omega(\boldsymbol{P}^2 + \boldsymbol{X}^2) \equiv -\frac{\omega}{2}\frac{\partial^2}{\partial x^2} + \frac{\omega}{2}x^2$$

The Hamilton ordinary differential equations are replaced by the Schrödinger equation, $\frac{d}{dt}|\psi\rangle = -\imath\frac{\boldsymbol{H}}{\hbar}|\psi\rangle$, a partial differential equation defining $\psi(x,t)$ from its initial condition

$$(\psi(x,0))_{x\in\mathbb{R}} : \imath\frac{\partial\psi}{\partial t}(x,t) = -\frac{\omega}{2}\frac{\partial^2\psi}{\partial x^2}(x,t) + \frac{\omega}{2}x^2\psi(x,t), \quad x \in \mathbb{R}$$

The average position reads

$$\langle \boldsymbol{X} \rangle_t = \langle \psi | \boldsymbol{X} | \psi \rangle = \tfrac{1}{\sqrt{2}} \int_{-\infty}^{+\infty} x|\psi|^2 dx.$$

The average impulsion reads

$$\langle \boldsymbol{P} \rangle_t = \langle \psi | \boldsymbol{P} | \psi \rangle = -\tfrac{\imath}{\sqrt{2}} \int_{-\infty}^{+\infty} \psi^* \frac{\partial\psi}{\partial x} dx,$$

(real quantity via an integration by part).

It is very convenient to introduced the annihilation operator $\boldsymbol{a}$ and creation operator $\boldsymbol{a}^\dagger$:

$$\boldsymbol{a} = \boldsymbol{X} + \imath \boldsymbol{P} \equiv \frac{1}{\sqrt{2}}\left(x + \frac{\partial}{\partial x}\right), \quad \boldsymbol{a}^\dagger = \boldsymbol{X} - \imath \boldsymbol{P} \equiv \tfrac{1}{\sqrt{2}}\left(x - \frac{\partial}{\partial x}\right).$$

We have

$$[\boldsymbol{X}, \boldsymbol{P}] = \tfrac{\imath}{2}\boldsymbol{I}, \quad [\boldsymbol{a}, \boldsymbol{a}^\dagger] = \boldsymbol{I}, \quad \boldsymbol{H} = \omega(\boldsymbol{P}^2 + \boldsymbol{X}^2) = \omega\left(\boldsymbol{a}^\dagger\boldsymbol{a} + \tfrac{1}{2}\boldsymbol{I}\right)$$

where $\boldsymbol{I}$ stands for the identity operator.

Since $[\boldsymbol{a}, \boldsymbol{a}^\dagger] = \boldsymbol{I}$, the spectral decomposition of $\boldsymbol{a}^\dagger\boldsymbol{a}$ is simple. The Hermitian operator $\boldsymbol{N} = \boldsymbol{a}^\dagger\boldsymbol{a}$, the photon-number operator, admits $\mathbb{N}$ as non degenerate spectrum. The normalized eigenstate $|n\rangle$ associated to $n \in \mathbb{N}$, is denoted by $|n\rangle$. Thus the underlying Hilbert space reads

$$\mathcal{H} = \left\{ \sum_{n\geq 0} \psi_n |n\rangle,\ (\psi_n)_{n\geq 0} \in l^2(\mathbb{C}) \right\}$$

where $(|n\rangle)_{n\in\mathbb{N}}$ is the Hilbert basis of photon-number states (also called Fock states). For $n > 0$, we have

$$\boldsymbol{a}|n\rangle = \sqrt{n}\,|n-1\rangle, \quad \boldsymbol{a}^\dagger|n\rangle = \sqrt{n+1}\,|n+1\rangle.$$

The ground state $|0\rangle$ is characterized by $\boldsymbol{a}|0\rangle = 0$. It corresponds to the Gaussian function $\psi_0(x) = \frac{1}{\pi^{1/4}}\exp(-x^2/2)$.

For any function f we have the following commutations

$$\boldsymbol{a} f(\boldsymbol{N}) = f(\boldsymbol{N} + \boldsymbol{I})\boldsymbol{a}, \quad \boldsymbol{a}^\dagger f(\boldsymbol{N}) = f(\boldsymbol{N} - \boldsymbol{I})\boldsymbol{a}^\dagger.$$

In particular for any angle θ, $e^{i\theta\boldsymbol{N}}\boldsymbol{a}e^{-i\theta\boldsymbol{N}} = e^{-i\theta}\boldsymbol{a}$.

For any amplitude $\alpha \in \mathbb{C}$, the Glauber displacement unitary operator $\boldsymbol{D}_\alpha$ is defined by

$$\boldsymbol{D}_\alpha = e^{\alpha\,\boldsymbol{a}^\dagger - \alpha^*\boldsymbol{a}}$$

We have $\boldsymbol{D}_\alpha^{-1} = \boldsymbol{D}_\alpha^\dagger = \boldsymbol{D}_{-\alpha}$. The following Glauber formula is useful: if two operators $\boldsymbol{A}$ and $\boldsymbol{B}$ commute with their commutator, i.e., if $[\boldsymbol{A}, [\boldsymbol{A}, \boldsymbol{B}]] = [\boldsymbol{B}, [\boldsymbol{A}, \boldsymbol{B}]] = 0$, then we have $e^{\boldsymbol{A}+\boldsymbol{B}} = e^{\boldsymbol{A}}\,e^{\boldsymbol{B}}\,e^{-\frac{1}{2}[\boldsymbol{A},\boldsymbol{B}]}$. Since $\boldsymbol{A} = \alpha\boldsymbol{a}^\dagger$ and $\boldsymbol{B} = -\alpha^*\boldsymbol{a}$ are in this case, we have another expression for $\boldsymbol{D}_\alpha$

$$\boldsymbol{D}_\alpha = e^{-\frac{|\alpha|^2}{2}}\,e^{\alpha\boldsymbol{a}^\dagger}e^{-\alpha^*\boldsymbol{a}} = e^{+\frac{|\alpha|^2}{2}}\,e^{-\alpha^*\boldsymbol{a}}e^{\alpha\boldsymbol{a}^\dagger}.$$

The terminology displacement has its origin in the following property derived from Baker-Campbell-Hausdorff formula:

$$\forall\alpha \in \mathbb{C}, \quad \boldsymbol{D}_{-\alpha}\boldsymbol{a}\boldsymbol{D}_\alpha = \boldsymbol{a} + \alpha \quad \text{and} \quad \boldsymbol{D}_{-\alpha}\boldsymbol{a}^\dagger\boldsymbol{D}_\alpha = \boldsymbol{a}^\dagger + \alpha^*.$$

To the classical state (x, p) is associated a quantum state usually called coherent state of complex amplitude $\alpha = (x + \imath p)/\sqrt{2}$ and denoted by $|\alpha\rangle$:

$$|\alpha\rangle = \boldsymbol{D}_\alpha|0\rangle = e^{-\frac{|\alpha|^2}{2}}\sum_{n=0}^{+\infty}\tfrac{\alpha^n}{\sqrt{n!}}|n\rangle. \tag{A.1}$$

$|\alpha\rangle$ corresponds to the translation of the Gaussian profile corresponding to vacuum state $|0\rangle$:

$$|\alpha\rangle \equiv \left(\mathbb{R} \ni x \mapsto \tfrac{1}{\pi^{1/4}} e^{\imath\sqrt{2}x\Im\alpha} e^{-\frac{(x-\sqrt{2}\Re\alpha)^2}{2}} \right).$$

This usual notation is potentially ambiguous: the coherent state $|\alpha\rangle$ is very different from the photon-number state $|n\rangle$ where n is a non negative integer: The probability p_n to obtain $n \in \mathbb{N}$ during the measurement of $\boldsymbol{N}$ with $|\alpha\rangle$ obeys to a Poisson law $p_n = e^{-|\alpha|^2}|\alpha|^{2n}/n!$. The resulting average energy is thus given by $\langle\alpha|\boldsymbol{N}|\alpha\rangle = |\alpha|^2$. Only for $\alpha = 0$ and $n = 0$, these quantum states coincide.

The coherent state $\alpha \in \mathbb{C}$ is the unitary eigenstate of $\boldsymbol{a}$ associated to the eigenvalue $\alpha \in \mathbb{C}$: $\boldsymbol{a}|\alpha\rangle = \alpha|\alpha\rangle$. Since $\boldsymbol{H}/\hbar = \omega(\boldsymbol{N} + \frac{1}{2})$, the solution of the Schrödinger equation $\frac{d}{dt}|\psi\rangle = -\imath\frac{\boldsymbol{H}}{\hbar}|\psi\rangle$, with initial value a coherent state $|\psi\rangle_{t=0} = |\alpha_0\rangle$ ($\alpha_0 \in \mathbb{C}$) remains a coherent state with time varying amplitude $\alpha_t = e^{-\imath\omega t}\alpha_0$:

$$|\psi\rangle_t = e^{-\imath\omega t/2}|\alpha_t\rangle.$$

These coherent solutions are the quantum counterpart of the classical solutions: $x_t = \sqrt{2}\Re(\alpha_t)$ and $p_t = \sqrt{2}\Im(\alpha_t)$ are solutions of the classical Hamilton equations $\frac{d}{dt}x = \omega p$ and $\frac{d}{dt}p = -\omega x$ since $\frac{d}{dt}\alpha_t = -\imath\omega\alpha_t$. The addition of a control input, a classical drive of amplitude $u \in \mathbb{R}$, yields to the following control Schrödinger equation

$$\frac{d}{dt}|\psi\rangle = -\imath\Big(\omega\left(\boldsymbol{a}^\dagger\boldsymbol{a} + \tfrac{1}{2}\right) + u(\boldsymbol{a} + \boldsymbol{a}^\dagger)\Big)|\psi\rangle$$

It is the quantum version of the control classical harmonic oscillator

$$\frac{d}{dt}x = \omega p, \quad \frac{d}{dt}p = -\omega x - u\sqrt{2}.$$

A possible definition of the Wigner function W^ρ attached to any density operator ρ is as follows:

$$W^\rho : \mathbb{C} \ni \alpha \to \tfrac{2}{\pi}\,\mathrm{Tr}\left(e^{i\pi\boldsymbol{N}}e^{-\alpha\boldsymbol{a}^\dagger + \alpha^*\boldsymbol{a}}\rho e^{\alpha\boldsymbol{a}^\dagger - \alpha^*\boldsymbol{a}}\right) \in [-2/\pi, 2/\pi]$$

where $\alpha = x + ip$ is a position in the phase-plane (x, p) of the classical oscillator. With the correspondences

$$\begin{aligned}
&\frac{\partial}{\partial\alpha} = \tfrac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial p}\right), \quad \frac{\partial}{\partial\alpha^*} = \tfrac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial p}\right)\\
&W^{\rho\boldsymbol{a}} = \left(\alpha - \tfrac{1}{2}\frac{\partial}{\partial\alpha^*}\right)W^\rho, \quad W^{\boldsymbol{a}\rho} = \left(\alpha + \tfrac{1}{2}\frac{\partial}{\partial\alpha^*}\right)W^\rho \qquad \text{(A.2)}\\
&W^{\rho\boldsymbol{a}^\dagger} = \left(\alpha^* + \tfrac{1}{2}\frac{\partial}{\partial\alpha}\right)W^\rho, \quad W^{\boldsymbol{a}^\dagger\rho} = \left(\alpha^* - \tfrac{1}{2}\frac{\partial}{\partial\alpha}\right)W^\rho
\end{aligned}$$

the Lindblad-Kossakovski governing the evolution of the density operator ρ of a quantum oscillator, with damping time constant $1/\kappa > 0$ and resonant drive of real amplitude u,

$$\frac{d}{dt}\rho = u[\boldsymbol{a}^\dagger - \boldsymbol{a}, \rho] + \kappa\left(\boldsymbol{a}\rho\boldsymbol{a}^\dagger - (\boldsymbol{N}\rho + \rho\boldsymbol{N})/2\right),$$

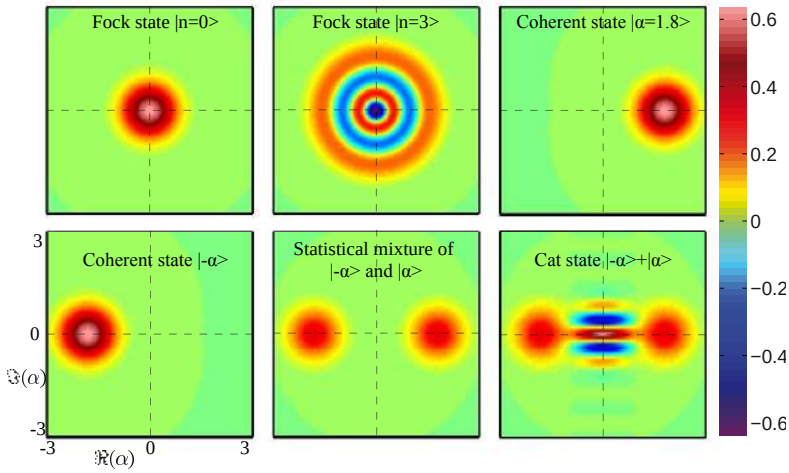


Figure A.1. Wigner function of typical quantum states of an harmonic oscillator.

becomes a convection-diffusion equation for the Wigner function W^ρ

$$\frac{\partial W^\rho}{\partial t} = \tfrac{\kappa}{2}\left(\frac{\partial}{\partial x}\Big((x-\bar\alpha)W^\rho\Big) + \frac{\partial}{\partial p}\Big(pW^\rho\Big) + \tfrac{1}{4}\Delta W^\rho\right)$$

where Δ denotes the Laplacian operator $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial p^2}$. The solutions converge toward the Gaussian steady-state $W^{\overline\rho}(x,p) = \frac{2}{\pi}e^{-2(x-\bar\alpha)^2-2p^2}$, where $\overline\rho = |\bar\alpha\rangle\langle\bar\alpha|$ is the coherent state of amplitude $\bar\alpha = 2u/\kappa$.

B. Qubit

The underlying Hilbert space $\mathcal{H} = \mathbb{C}^2 = \{c_g|g\rangle + c_e|e\rangle,\ c_g, c_e \in \mathbb{C}\}$ where $(|g\rangle, |e\rangle)$ is the ortho-normal frame formed by the ground state $|g\rangle$ and the excited state $|e\rangle$. It is usual to consider the following operators on $\mathcal{H}$:

$$\begin{aligned}&\boldsymbol{\sigma_-} = |g\rangle\langle e|, \quad \boldsymbol{\sigma_+} = \boldsymbol{\sigma_-}^\dagger = |e\rangle\langle g|, \quad \boldsymbol{\sigma_x} = \boldsymbol{\sigma_-} + \boldsymbol{\sigma_+} = |g\rangle\langle e| + |e\rangle\langle g|,\\ &\boldsymbol{\sigma_y} = i\boldsymbol{\sigma_-} - i\boldsymbol{\sigma_+} = i|g\rangle\langle e| - i|e\rangle\langle g|, \quad \boldsymbol{\sigma_z} = \boldsymbol{\sigma_+}\boldsymbol{\sigma_-} - \boldsymbol{\sigma_-}\boldsymbol{\sigma_+} = |e\rangle\langle e| - |g\rangle\langle g|.\end{aligned} \tag{B.1}$$

$\boldsymbol{\sigma_x}$, $\boldsymbol{\sigma_y}$ and $\boldsymbol{\sigma_z}$ are the Pauli operators. They are square root of $\boldsymbol{I}$: $\boldsymbol{\sigma_x}^2 = \boldsymbol{\sigma_y}^2 = \boldsymbol{\sigma_z}^2 = \boldsymbol{I}$. They anti-commute

$$\boldsymbol{\sigma_x}\boldsymbol{\sigma_y} = -\boldsymbol{\sigma_y}\boldsymbol{\sigma_x} = i\boldsymbol{\sigma_z}, \quad \boldsymbol{\sigma_y}\boldsymbol{\sigma_z} = -\boldsymbol{\sigma_z}\boldsymbol{\sigma_y} = i\boldsymbol{\sigma_x}, \quad \boldsymbol{\sigma_z}\boldsymbol{\sigma_x} = -\boldsymbol{\sigma_x}\boldsymbol{\sigma_z} = i\boldsymbol{\sigma_y}$$

and thus $[\boldsymbol{\sigma_x}, \boldsymbol{\sigma_y}] = 2i\boldsymbol{\sigma_z}$, $[\boldsymbol{\sigma_y}, \boldsymbol{\sigma_z}] = 2i\boldsymbol{\sigma_x}$, $[\boldsymbol{\sigma_z}, \boldsymbol{\sigma_x}] = 2i\boldsymbol{\sigma_y}$. The uncontrolled evolution is governed by the Hamiltonian $\boldsymbol{H}/\hbar = \omega\boldsymbol{\sigma_z}/2$ where $\omega > 0$ is the qubit pulsation. Thus the solution of $\frac{d}{dt}|\psi\rangle = -i\frac{\boldsymbol{H}}{\hbar}|\psi\rangle$ is given by

$$|\psi\rangle_t = e^{-i\left(\frac{\omega t}{2}\right)\boldsymbol{\sigma_z}}|\psi\rangle_0 = \cos\left(\tfrac{\omega t}{2}\right)|\psi\rangle_0 - i\sin\left(\tfrac{\omega t}{2}\right)\boldsymbol{\sigma_z}|\psi\rangle_0$$

since for any angle θ we have

$$e^{i\theta\boldsymbol{\sigma_x}} = \cos\theta + i\sin\theta\boldsymbol{\sigma_x}, \quad e^{i\theta\boldsymbol{\sigma_y}} = \cos\theta + i\sin\theta\boldsymbol{\sigma_y}, \quad e^{i\theta\boldsymbol{\sigma_z}} = \cos\theta + i\sin\theta\boldsymbol{\sigma_z}.$$

Since the Pauli operators anti-commute, we have the useful relationships:

$$e^{i\theta\boldsymbol{\sigma_x}}\boldsymbol{\sigma_y} = \boldsymbol{\sigma_y}e^{-i\theta\boldsymbol{\sigma_x}}, \quad e^{i\theta\boldsymbol{\sigma_y}}\boldsymbol{\sigma_z} = \boldsymbol{\sigma_z}e^{-i\theta\boldsymbol{\sigma_y}}, \quad e^{i\theta\boldsymbol{\sigma_z}}\boldsymbol{\sigma_x} = \boldsymbol{\sigma_x}e^{-i\theta\boldsymbol{\sigma_z}}.$$

The orthogonal projector $\rho = |\psi\rangle\langle\psi|$, the density operator associated to the pure state $|\psi\rangle$, obeys to the Liouville equation $\frac{d}{dt}\rho = -\frac{i}{\hbar}[\boldsymbol{H}, \rho]$. Mixed quantum states are described by ρ that are Hermitian, non-negative and of trace one. For a qubit, the Bloch sphere representation is a useful tool exploiting the smooth correspondence between such ρ and the unit ball of $\mathbb{R}^3$ considered as Euclidian space:

$$\rho = \frac{\boldsymbol{I} + x\boldsymbol{\sigma_x} + y\boldsymbol{\sigma_y} + z\boldsymbol{\sigma_z}}{2}, \quad (x, y, z) \in \mathbb{R}^3, \quad x^2 + y^2 + z^2 \leq 1.$$

$(x, y, z) \in \mathbb{R}^3$ are the coordinates in the orthonormal frame $(\vec{\imath}, \vec{\jmath}, \vec{k})$ of the Bloch vector $\vec{M} \in \mathbb{R}^3$. This vector lives on or inside the unit sphere, called Bloch sphere:

$$\vec{M} = x\vec{\imath} + y\vec{\jmath} + z\vec{k}.$$

Since $\mathrm{Tr}\left(\rho^2\right) = x^2 + y^2 + z^2$, $\vec{M}$ is on the Bloch sphere when ρ is of rank one and thus is a pure state. The translation of Liouville equation on $\vec{M}$ yields with $\boldsymbol{H}/\hbar = \omega\boldsymbol{\sigma_z}/2$: $\frac{d}{dt}\vec{M} = \omega\vec{k} \times \vec{M}$. For the two-level system with the coherent drive described by the complex valued control u, $\boldsymbol{H}/\hbar = \frac{\omega}{2}\boldsymbol{\sigma_z} + \frac{\Re(u)}{2}\boldsymbol{\sigma_x} + \frac{\Im(u)}{2}\boldsymbol{\sigma_y}$ and the Liouville equation reads, with the Bloch vector $\vec{M}$ representation,

$$\frac{d}{dt}\vec{M} = (\Re(u)\vec{\imath} + \Im(u)\vec{\jmath} + \omega\vec{k}) \times \vec{M}.$$

C. Jaynes-Cumming Hamiltonians and propagators

The Jaynes-Cummings Hamiltonian [36] is the simplest Hamiltonian describing the interaction between an harmonic oscillator and a qubit. Such an interaction admits two regimes, the resonant one where the oscillator and the qubit exchange energy, the dispersive one where the oscillator pulsation depends on the qubit state and where the qubit pulsation, slightly different from the oscillator pulsation, depends on the number of vibration quanta. We recall below the simplest forms of these Hamiltonians in the interaction frame. A deeper and complete presentation can be found in [33].

The resonant Hamiltonian $\boldsymbol{H}_{res}$ is given by

$$\boldsymbol{H}_{res}/\hbar = if(t)\left(\boldsymbol{a}^\dagger \otimes \boldsymbol{\sigma_-} - \boldsymbol{a} \otimes \boldsymbol{\sigma_+}\right) = if(t)\left(\boldsymbol{a}^\dagger \otimes |g\rangle\langle e| - \boldsymbol{a} \otimes |e\rangle\langle g|\right) \tag{C.1}$$

whereas the dispersive one $\boldsymbol{H}_{disp}$ is a simple tensor product:

$$\boldsymbol{H}_{disp}/\hbar = f(t)\,\boldsymbol{N} \otimes \boldsymbol{\sigma_z} = f(t)\,\boldsymbol{N} \otimes \left(|e\rangle\langle e| - |g\rangle\langle g|\right) \tag{C.2}$$

where $f(t)$ is a known real parameter depending possibly on the time t.

Simple computations show that the resonant propagator $\boldsymbol{U}_{res}$ between t_0 and t_1 associated to $\boldsymbol{H}_{res}$, i.e., the solution of Cauchy problem

$$\frac{d}{dt}\boldsymbol{U} = -i\frac{\boldsymbol{H}_{res}}{\hbar}\boldsymbol{U}, \quad U(t_0) = \boldsymbol{I},$$

is explicit and given by the following compact formulae:

$$\begin{aligned}\boldsymbol{U}_{res}(t_0,t_1) = \cos\left(\boldsymbol{N}\int_{t_0}^{t_1} f\right)\otimes|g\rangle\langle g| + \cos\left((\boldsymbol{N}+\boldsymbol{I})\int_{t_0}^{t_1} f\right)\otimes|e\rangle\langle e| \\ - \boldsymbol{a}\frac{\sin\left(\boldsymbol{N}\int_{t_0}^{t_1} f\right)}{\sqrt{\boldsymbol{N}}}\otimes|e\rangle\langle g| + \frac{\sin\left(\boldsymbol{N}\int_{t_0}^{t_1} f\right)}{\sqrt{\boldsymbol{N}}}\boldsymbol{a}^\dagger\otimes|g\rangle\langle e|. \quad \text{(C.3)}\end{aligned}$$

It is instructive to check that $\boldsymbol{U}_{res}^\dagger\boldsymbol{U}_{res} = \boldsymbol{I}$. Similarly, the dispersive propagator $\boldsymbol{U}_{disp}$ between t_1 and t_2 associated to $\boldsymbol{H}_{disp}$ is given by

$$\boldsymbol{U}_{disp}(t_0,t_1) = \exp\left(i\boldsymbol{N}\int_{t_0}^{t_1} f\right)\otimes|g\rangle\langle g| + \exp\left(-i\boldsymbol{N}\int_{t_0}^{t_1} f\right)\otimes|e\rangle\langle e|. \quad \text{(C.4)}$$

References

[1] C. Altafini and F. Ticozzi, *Modeling and control of quantum systems: An introduction*, Automatic Control, IEEE Transactions on, **57**(8) (2012), 1898–1917.

[2] H Amini, C. Pellegrini, and P. Rouchon, *Stability of continuous-time quantum filters with measurement imperfections*, arXiv:1312.0418v1, 2013.

[3] H. Amini, R.A. Somaraju, I. Dotsenko, C. Sayrin, M. Mirrahimi, and P. Rouchon, *Feedback stabilization of discrete-time quantum systems subject to non-demolition measurements with imperfections and delays*, Automatica, **49**(9) (September 2013), 2683–2692.

[4] S. Attal, A. Joye, and C.-A. Pillet, editors, *Open Quantum Systems III: Recent Developments*. Lecture notes in Mathematics 1880, Springer, 2006.

[5] A. Barchielli and M. Gregoratti, *Quantum Trajectories and Measurements in Continuous Time: the Diffusive Case*, Springer Verlag, 2009.

[6] S. M. Barnett and P. M. Radmore, *Methods in Theoretical Quantum Optics*, Oxford University Press, 2003.

[7] L. Baudoin and J. Salomon, *Constructive solution of a bilinear optimal control problem for a Schrödinger equation*, Systems and Control Letters, **57** (2008), 453–464.

[8] L. Baudouin, O. Kavian, and J.P. Puel, *Regularity for a Schrödinger equation with singular potentials and application to bilinear optimal control*, J. Differential Equations, **216** (2005), 188–222.

[9] M; Bauer, T. Benoist, and D. Bernard:, *Repeated quantum non-demolition measurements: Convergence and continuous time limit*, Ann. Henri Poincare, **14** (2013), 639–679.

[10] K. Beauchard and J.-M. Coron, *Controllability of a quantum particle in a moving potential well*, J. of Functional Analysis, **232** (2006), 328–389.

[11] K. Beauchard, J.-M. Coron, and P. Rouchon, *Controllability issues for continuous spectrum systems and ensemble controllability of Bloch equations*, Communications in Mathematical Physics, **296** (2010), 525–557.

[12] K. Beauchard, P.S. Pereira da Silva, and P. Rouchon, *Stabilization of an arbitrary profile for an ensemble of half-spin systems*, Automatica, **49**(7) (July 2013), 2133–2137.

[13] V.P. Belavkin, *Quantum stochastic calculus and quantum nonlinear filtering*, Journal of Multivariate Analysis, **42**(2) (1992), 171–201.

[14] G. Birkhoff, *Extensions of Jentzch's theorem*, Trans. Amer. Math. Soc., **85** (1957), 219–227.

[15] B. Bonnard, O. Cots, S.J. Glaser, M. Lapert, D. Sugny, and Yun Zhang, *Geometric optimal control of the contrast imaging problem in nuclear magnetic resonance*, Automatic Control, IEEE Transactions on, **57**(8) (2012), 1957–1969.

[16] L. Bouten and R. van Handel, *Quantum Stochastics and Information: Statistics, Filtering and Control*, chapter On the separation principle of quantum control. World Scientific, 2008.

[17] P. Campagne-Ibarcq, E. Flurin, N. Roch, D. Darson, P. Morfin, M. Mirrahimi, M. H. Devoret, F. Mallet, and B. Huard, *Persistent control of a superconducting qubit by stroboscopic measurement feedback*, Phys. Rev. X, **3**(2) (May 2013), 021008–.

[18] H. Carmichael, *An Open Systems Approach to Quantum Optics*, Springer-Verlag, 1993.

[19] _____, *Statistical Methods in Quantum Optics 1: Master Equations and Fokker-Planck Equations*, Springer, 1999.

[20] H. Carmichael, *Statistical Methods in Quantum Optics 2: Non-Classical Fields*, Spinger, 2007.

[21] T. Chambrion, P. Mason, M. Sigalotti, and M. Boscain, *Controllability of the discrete-spectrum Schrödinger equation driven by an external field*, Ann. Inst. H. Poincaré Anal. Non Linéaire, **26**(1) (2009), 329–349.

[22] A. A. Clerk, M. H. Devoret, S. M. Girvin, Florian Marquardt, and R. J. Schoelkopf, *Introduction to quantum noise, measurement, and amplification*, Rev. Mod. Phys., **82**(2) (April 2010), 1155–1208.

[23] D. D'Alessandro, *Introduction to Quantum Control and Dynamics*, Chapman & Hall/CRC, 2008.

[24] J. Dalibard, Y. Castion, and K. Mølmer, *Wave-function approach to dissipative processes in quantum optics*, Phys. Rev. Lett., **68**(5) (1992), 580–583.

[25] E.B. Davies, *Quantum Theory of Open Systems*, Academic Press, 1976.

[26] I. Dotsenko, M. Mirrahimi, M. Brune, S. Haroche, J.-M. Raimond, and P. Rouchon, *Quantum feedback by discrete quantum non-demolition measurements: towards on-demand generation of photon-number states*, Physical Review A, **80** (2009), 013805–013813.

[27] S. Ervedoza and J.-P. Puel, *Approximate controllability for a system of Schrödinger equations modeling a single trapped ion*, Annales de l'Institut Henri Poincare (C) Non Linear Analysis, **26**(6) (2009), 2111 – 2136.

[28] C.W. Gardiner and P. Zoller, *Quantum noise*, Springer, third edition, 2010.

[29] A. Garon, S. J. Glaser, and D. Sugny, *Time-optimal control of SU(2) quantum operations*, Phys. Rev. A, **88**(4) (October 2013), 043422–.

[30] J. Gough and M.R. James, *The series product and its application to quantum feedforward and feedback networks*, Automatic Control, IEEE Transactions on, **54**(11) (2009), 2530–2544.

[31] A. Grigoriu, H. Rabitz, and G. Turinici, *Controllability analysis of quantum systems immersed within an engineered environment*, Journal of Mathematical Chemistry, **51**(6) (2013), 1548–1560.

[32] R. Hamerly and H. Mabuchi, *Advantages of coherent feedback for cooling quantum oscillators*, Phys. Rev. Lett., **109**(17) (October 2012), 173602–.

[33] S. Haroche and J.M. Raimond, *Exploring the Quantum: Atoms, Cavities and Photons*, Oxford University Press, 2006.

[34] M.R. James, *Quantum feedback control*, In Control Conference (CCC), 2011 30th Chinese, pages 26–34, 2011.

[35] M.R. James, H.I. Nurdin, and I.R. Petersen, *H infinity control of linear quantum stochastic systems*, Automatic Control, IEEE Transactions on, **53**(8) (2008), 1787–1803.

[36] E.T. Jaynes and F.W. Cummings, *Comparison of quantum and semiclassical radiation theories with application to the beam maser*, Proceedings of the IEEE, **51**(1) (1963), 89–109.

[37] A. Kastler, *Optical methods for studying Hertzian resonances*, Science, **158**(3798) (October 1967), 214–221.

[38] A. Kossakowski, *On quantum statistical mechanics of non-Hamiltonian systems*, Reports on Mathematical Physics, 3, 1972.

[39] H.J. Kushner, *Introduction to Stochastic Control*, Holt, Rinehart and Wilson, INC., 1971.

[40] Z. Leghtas, *Quantum state engineering and stabilization*, PhD thesis, Mines ParisTech, 2012.

[41] G. Lindblad, *On the generators of quantum dynamical semigroups*, Communications in Mathematical Physics, 48, 1976.

[42] S. Lloyd, *Coherent quantum feedback*, Phys. Rev. A, **62**(2) (July 2000), 022108–.

[43] H. Mabuchi and N. Khaneja, *Principles and applications of control in quantum systems*, International Journal of Robust and Nonlinear Control, **15**(15) (2005), 647–667.

[44] J.C Maxwell, *On governors*, Proc. Roy. Soc. (London), **16** (1868).

[45] M. Mirrahimi, Z. Leghtas, V.V. Albert, S. Touzard, R.J.. Schoelkopf, L. Jiang, and M.H. Devoret, *Dynamically protected cat-qubits: a new paradigm for universal quantum computation*, to appear in New Journal of Physics arXiv:1312.2017v1, 2014.

[46] M. Mirrahimi and R. Van Handel, *Stabilizing feedback controls for quantum systems*, SIAM Journal on Control and Optimization, **46**(2) (2007), 445–467.

[47] M.A. Nielsen and I.L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, 2000.

[48] N.C. Nielsen, C. Kehlet, S.J. Glaser, and N. Khaneja, *Optimal control methods in nmr spectroscopy*, In Encyclopedia of Nuclear Magnetic Resonance, pages –. John Wiley & Sons, Ltd, 2010.

[49] D. Petz, *Monotone metrics on matrix spaces*, Linear Algebra and its Applications, **244** (1996), 81–96.

[50] J. F. Poyatos, J. I. Cirac, and P. Zoller, *Quantum reservoir engineering with laser cooled trapped ions*, Phys. Rev. Lett., **77**(23) (December 1996), 4728–4731.

[51] D. Reeb, M. J. Kastoryano, and M. M. Wolf, *Hilbert's projective metric in quantum information theory*, Journal of Mathematical Physics, **52**(8) (August 2011), 082201.

[52] P. Rouchon, *Fidelity is a sub-martingale for discrete-time quantum filters*, IEEE Transactions on Automatic Control, **56**(11) (2011), 2743–2747.

[53] S. Sarlette, M. Brune, J.M. Raimond, and P. Rouchon, *Stabilization of nonclassical states of the radiation field in a cavity by reservoir engineering*, Phys. Rev. Lett., **107** (2011), 010402.

[54] S. Sarlette, Z. Leghtas, M. Brune, J.M. Raimond, and P. Rouchon, *Stabilization of nonclassical states of one and two-mode radiation fields by reservoir engineering*, Phys. Rev. A, **86** (2012), 012114.

[55] C. Sayrin, *Préparation et stabilisation d'un champ non classique en cavité par rétroaction quantique*, PhD thesis, Université Paris VI, 2011.

[56] C. Sayrin, I. Dotsenko, X. Zhou, B. Peaudecerf, Th. Rybarczyk, S. Gleyzes, P. Rouchon, M. Mirrahimi, H. Amini, M. Brune, J.M. Raimond, and S. Haroche, *Real-time quantum feedback prepares and stabilizes photon number states*, Nature, **477** (2011), 73–77.

[57] R. Sepulchre, A. Sarlette, and P. Rouchon, *Consensus in non-commutative spaces*, In Decision and Control (CDC), 2010 49th IEEE Conference on, pages 6596–6601, 2010.

[58] A. Somaraju, I. Dotsenko, C. Sayrin, and P. Rouchon, *Design and stability of discrete-time quantum filters with measurement imperfections*, In American Control Conference, pages 5084–5089, 2012.

[59] A. Somaraju, M. Mirrahimi, and P Rouchon, *Approximate stabilization of an infinite dimensional quantum stochastic system*, Rev. Math. Phys., **25**(01) (January 2013), 1350001–.

[60] R. van Handel, *Filtering, Stability, and Robustness*, PhD thesis, California Institute of Technology, 2007.

[61] _____, *The stability of quantum Markov filters*, Infin. Dimens. Anal. Quantum Probab. Relat. Top., **12** (2009), 153–172.

[62] R. Vijay, C. Macklin, D. H. Slichter, S. J. Weber, K. W. Murch, R. Naik, A. N. Korotkov, and I. Siddiqi, *Stabilizing Rabi oscillations in a superconducting qubit using quantum feedback*, Nature, **490**(7418) (2012), 77–80.

[63] H.M. Wiseman and G.J. Milburn, *Quantum Measurement and Control*, Cambridge University Press, 2009.

[64] X. Zhou, I. Dotsenko, B. Peaudecerf, T. Rybarczyk, C. Sayrin, J.M. Raimond S. Gleyzes, M. Brune, and S. Haroche, *Field locked to Fock state by quantum feedback with single photon corrections*, Physical Review Letter, **108** (2012), 243602.

Centre Automatique et Systèmes, Mines ParisTech, 60, Bd Saint-Michel, 75006 Paris, France
E-mail: pierre.rouchon@mines-paristech.fr

Time-inconsistent optimal control problems

Jiongmin Yong

Abstract. An optimal control problem is time-consistent if for any initial pair of time and state, whenever there exists an optimal control, it will stay optimal thereafter. In real world, however, such kind of time-consistency is hardly true, mainly due to the time-inconsistency of decision maker's time-preference and/or risk-preference. In another word, most optimal control problems, if not all, are not time-consistent, or time-inconsistent. In this paper, some general time-inconsistent optimal control problems are formulated for stochastic differential equations. Recent works of the author concerning the (time-consistent) equilibrium solutions to the time-inconsistent problems are surveyed.

Mathematics Subject Classification (2010). Primary 93E20, 49L20, 49N05; Secondary 49N70.

Keywords. Stochastic optimal control, time inconsistency, equilibrium solution, Hamilton-Jacobi-Bellman equation, differential games, linear-quadratic problem, Riccati equation.

1. Introduction

Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a complete filtered probability space on which a d-dimensional standard Brownian motion $W(\cdot)$ is defined, whose natural filtration is $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ (augmented by all the $\mathbb{P}$-null sets). Let $T > 0$. For any $t \in [0, T)$, we consider the following controlled stochastic differential equation (SDE, for short):

$$\begin{cases} dX(s) = b(s, X(s), u(s))ds + \sigma(s, X(s), u(s))dW(s), & s \in [t, T], \\ X(t) = x, \end{cases} \tag{1.1}$$

where $b : [0, T] \times \mathbb{R}^n \times U \to \mathbb{R}^n$ and $\sigma : [0, T] \times \mathbb{R}^n \times U \to \mathbb{R}^{n\times d}$ are suitable deterministic maps with U being a metric space. In the above, $X : [0, T] \times \Omega \to \mathbb{R}^n$ is called a *state process*, $u : [t, T] \times \Omega \to U$ is called a *control process*, and $(t, x) \in \mathcal{D}^p$ is called an *initial pair*, where $p > 1$, and

$$\mathcal{D}^p = \Big\{(t, x) \;\Big|\; t \text{ is an } \mathbb{F}\text{-stopping time valued in } [0, T), \\ \text{and } x \text{ is } \mathcal{F}_t\text{-measurable, } \mathbb{R}^n\text{-valued, with } \mathbb{E}|x|^p < \infty\Big\}.$$

We define the set of all *admissible control processes* by the following:

$$\mathcal{U}[t, T] = \Big\{u : [t, T] \times \Omega \to U \;\Big|\; u(\cdot) \text{ is } \mathbb{F}\text{-progressively measurable}\Big\}. \tag{1.2}$$

Under some mild conditions, for any $(t, x) \in \mathcal{D}^p$ and $u(\cdot) \in \mathcal{U}[t, T]$, (1.1) admits a unique strong solution $X(\cdot) \equiv X(\cdot\,; t, x, u(\cdot))$. To measure the performance of the control process

Proceedings of the International Congress of Mathematicians, Seoul, 2014

$u(\cdot)\in\mathcal{U}[t,T]$, we introduce the following cost functional

$$J^0(t,x;u(\cdot))=\mathbb{E}_t\Big[\int_t^T e^{-\delta(s-t)}g^0(s,X(s),u(s))ds+e^{-\delta(T-t)}h^0(X(T))\Big],\qquad(1.3)$$

with some constant $\delta\geq 0$ (called the *discount rate*), some maps $g^0:[0,T]\times\mathbb{R}^n\times U\to\mathbb{R}$ and $h^0:\mathbb{R}^n\to\mathbb{R}$, and $\mathbb{E}_t[\,\cdot\,]=\mathbb{E}[\,\cdot\,|\,\mathcal{F}_t]$. On the right hand side of (1.3), the first term is referred to as a *running cost* and the second term is referred to as a *terminal cost*. We can pose the following optimal control problem.

Problem (C). For any $(t,x)\in\mathcal{D}^p$, find a $\bar u(\cdot)\in\mathcal{U}[t,T]$ such that

$$J^0(t,x;\bar u(\cdot))=\inf_{u(\cdot)\in\mathcal{U}[t,T]}J^0(t,x;u(\cdot))\equiv V^0(t,x).\qquad(1.4)$$

Any $\bar u(\cdot)\in\mathcal{U}[t,T]$ satisfying (1.4) is called an *optimal control* of Problem (C) for the initial pair (t,x), the corresponding state process $\bar X(\cdot)\equiv X(\cdot\,;t,x,\bar u(\cdot))$ and the pair $(\bar X(\cdot),\bar u(\cdot))$ are called the corresponding *optimal state process* and *optimal pair*, respectively. The function $V^0(\cdot\,,\cdot)$ defined by (1.4) is called the *value function* of Problem (C). For Problem (C), we have the following Bellman's principle of optimality (see [27, 28], also, [12, 39]): For any $\tau\in[t,T]$,

$$\begin{aligned}V^0(t,x)=\inf_{u(\cdot)\in\mathcal{U}[t,\tau]}\mathbb{E}_t\Big[&\int_t^\tau e^{-\delta(s-t)}g^0(s,X(s),u(s))ds\\&+e^{-\delta(\tau-t)}V^0\big(\tau,X(\tau;t,x,u(\cdot))\big)\Big],\end{aligned}\qquad(1.5)$$

where $\mathcal{U}[t,\tau]$ is defined similar to $\mathcal{U}[t,T]$, replacing $[t,T]$ by $[t,\tau]$ (see (1.2)). Now, if $(\bar X(\cdot),\bar u(\cdot))$ is an optimal pair of Problem (C) for the initial pair $(t,x)\in[0,T)\times\mathbb{R}^n$, then from (1.5), we obtain

$$\begin{aligned}&V^0(t,x)\\&=J^0(t,x;\bar u(\cdot))\\&=\mathbb{E}_t\Big[\int_t^\tau e^{-\delta(s-t)}g^0(s,\bar X(s),\bar u(s))ds+e^{-\delta(\tau-t)}J^0\big(\tau,\bar X(\tau;t,x,\bar u(\cdot));\bar u(\cdot)\big|_{[\tau,T]}\big)\Big]\\&\geqslant\inf_{u(\cdot)\in\mathcal{U}[t,\tau]}\mathbb{E}_t\Big[\int_t^\tau e^{-\delta(s-t)}g^0(s,X(s),u(s))ds+e^{-\delta(\tau-t)}V^0\big(\tau,X(\tau;t,x,u(\cdot))\big)\Big]\\&=V^0(t,x).\end{aligned}$$

Thus, one must have

$$\mathbb{E}_t\Big[J^0\big(\tau,\bar X(\tau);\bar u(\cdot)\big|_{[\tau,T]}\big)-V^0(\tau,\bar X(\tau))\Big]=0,\qquad\text{a.s.}$$

It follows that

$$J^0\big(\tau,\bar X(\tau);\bar u(\cdot)\big|_{[\tau,T]}\big)=V^0(\tau,\bar X(\tau))=\inf_{u(\cdot)\in\mathcal{U}[\tau,T]}J^0\big(\tau,\bar X(\tau);u(\cdot)\big),\quad\text{a.s.}$$

This means that the restriction $\bar u(\cdot)\big|_{[\tau,T]}\in\mathcal{U}[\tau,T]$ of an optimal control $\bar u(\cdot)\in\mathcal{U}[t,T]$ for the initial pair (t,x) on any later time interval $[\tau,T]$ is optimal for the initial pair

$\big(\tau, \bar{X}(\tau; t, x, \bar{u}(\cdot))\big) \in \mathcal{D}^p$. Such a phenomenon is called the *time-consistency* of Problem (C).

The advantage of the time-consistency is that for any given initial pair (t, x), if an optimal control $\bar{u}(\cdot)$ can be constructed for that (initial pair), then it will stay optimal thereafter (for the later initial pair along the optimal trajectory). However, in reality, the situation is rarely such ideal, namely, many real problems do not have the time-consistency. This is mainly caused by the time-inconsistency of people's *time-preferences* and/or *risk-preferences*, which we now briefly explain.

From common experience, we know that people usually over-discount the payoff/cost of the immediate future time period than that of the farther future time periods. It was pointed out by Hume ([16]) that passion is dominating over reason during the immediate future period. This is referred to as people's *time-preferences*. Mathematically, one can describe such a situation by what we call the *general discounting* which includes the so-called *quasi-exponential discounting*, *hyperbolic discounting*, and/or *non-exponential discounting* situations. This amounts to replacing the discount functions $e^{-\delta(s-t)}$ and $e^{-\delta(T-t)}$ appeared in (1.3) by some general functions $\lambda(s, t)$ and $\nu(T, t)$, or even more generally, we may consider the following cost functional:

$$J(t, x; u(\cdot)) = \mathbb{E}_t\Big[\int_t^T g(t, s, X(s), u(s))ds + h(t, X(T))\Big], \tag{1.6}$$

where the maps $g(\cdot)$ and $h(\cdot)$ explicitly depend on the initial time t in some general way. The optimal control problem associated with (1.1) and (1.6) will not be time-consistent, or *time-inconsistent*, in general, meaning that the restriction of an optimal control for a given initial pair on a later time interval might not be optimal for that corresponding initial pair.

On the other hand, different groups of people have different attitudes towards risks. One may be risk-averse (when making decisions for investment, buying insurance, etc.), or risk-seeking (when buying a lottery, gambling, etc.). These are referred to as people's *risk-preferences*. A classical way of describing people's risk-references is to use the so-called *expected utility* which can be traced back to Bernoulli's resolution of St.Petersburg's paradox ([2]). For a general expected utility theory, see the book by von Neumann–Morgenstern [25]. Later it was extended to the so-called *stochastic differential utility* ([8]) which can be represented by the adapted solutions to certain backward stochastic differential equations (BSDEs, for short) (see [13]). However, the paradoxes of Allais [1] and Ellsberg [14] showed that expected utility might not completely represent people's risk-preferences. In fact, even earlier, it was already realized by some scholars that the probability involved in the classical expected utility theory should be subjective, instead of objective, see for example, [6, 31, 32], etc. To describe people's risk-preferences, one may use the so-called *Choquet expected utility*, by which we mean that in the standard expected utility theory, replace the usual expectation by the so-called *Choquet integral* ([5], see also [7]). A special case of that is the expected utility with respect to the so-called *distorted probability* ([7, 33]) which is widely used in the insurance related studies and the behavioral finance. See [17, 20], and so on, for relevant results. On the other hand, inspired by the mean-variance problems, people could represent their dynamic risk-preferences by conditional variance. An interesting motivation from optimal control relevant to this is as follows. Practically, one hopes that the optimal control $\bar{u}(\cdot)$ and/or optimal state trajectory $\bar{X}(\cdot)$ are not too random. To achieve this, one could include conditional variance of the state-control pair $\mathrm{var}_t[X(\cdot)]$ and/or $\mathrm{var}_t[u(\cdot)]$) in

the cost functional, where

$$\begin{aligned}\mathrm{var}_t[X(s)] &= \mathbb{E}_t\big|X(s)-\mathbb{E}_t[X(s)]\big|^2 = \mathbb{E}_t|X(s)|^2 - \big|\mathbb{E}_t[X(s)]\big|^2,\\ \mathrm{var}_t[u(s)] &= \mathbb{E}_t\big|u(s)-\mathbb{E}_t[u(s)]\big|^2 = \mathbb{E}_t|u(s)|^2 - \big|\mathbb{E}_t[u(s)]\big|^2.\end{aligned}$$

Note that in the above, $\mathbb{E}_t[X(s)]$ and $\mathbb{E}_t[u(s)]$ are nonlinearly appeared. Suggested by the above, we see that it is possible to consider the following cost functional:

$$J(t,x;u(\cdot)) = \mathbb{E}_t\Big[\int_t^T g\big(s,X(s),u(s),\mathbb{E}_t[X(s)],\mathbb{E}_t[u(s)]\big)ds + h\big(X(T),\mathbb{E}[X(T)]\big)\Big].$$

Some concrete examples can be cooked up to show that if in the cost functional, one either has non-exponential discounting, or nonlinear appearance of conditional expectation of the state and/or control, the corresponding optimal control problem will be time-inconsistent in general. See [36, 38] for some details.

A natural question is: *What can we do for the time-inconsistent optimal control problems?* It is desired that one can obtain the so-called time-consistent equilibrium solutions for time-inconsistent problems. In this paper, we will survey some results obtained by the author in the recent years. For some relevant works, see [3, 4, 9–11, 19, 23, 24, 26, 29, 30, 34, 35].

2. Problem with general discounting

In this section, we consider state equation (1.1) with deterministic coefficients and with the following cost functional:

$$J(t,x;u(\cdot)) = \mathbb{E}_t\Big[\int_t^T g(t,s,X(s),u(s))ds + h(t,X(T))\Big]. \tag{2.1}$$

Clearly, our cost functional covers the non-exponential/hyperbolic discounting situations.

2.1. The problem. In what follows, we let $T>0$ be a fixed time horizon, and $U\subseteq\mathbb{R}^m$ be a closed subset, which could be either bounded or unbounded (it is allowed that $U=\mathbb{R}^m$). We will use $K>0$ as a generic constant which can be different from line to line. Let $\mathbb{S}^n$ be the set of all $(n\times n)$ symmetric real matrices. Denote

$$D[0,T] = \Big\{(t,s)\in[0,T]^2 \;\Big|\; 0\leqslant t\leqslant s\leqslant T\Big\}.$$

Recall the definition of $\mathcal{U}[t,T]$ from Section 1 (see (1.2)). Further, for $q\geqslant 1$, let

$$\begin{aligned}\mathcal{U}^q[t,T] = \Big\{u:[t,T]\times\Omega\to U \;\Big|\; & u(\cdot)\text{ is }\mathbb{F}\text{-progressively measurable},\\ & \mathbb{E}\int_t^T |u(s)|^q ds<\infty\Big\}.\end{aligned}$$

Note that in the case U is bounded, for different $q\geqslant 1$, all the $\mathcal{U}^q[t,T]$ are the same as $\mathcal{U}[t,T]$. But, if U is unbounded, $\mathcal{U}^q[t,T]$ will be different for different $q\in[1,\infty)$. We introduce the following standing assumptions.

(H1) The maps $b : [0,T] \times \mathbb{R}^n \times U \to \mathbb{R}^n$, $\sigma : [0,T] \times \mathbb{R}^n \times U \to \mathbb{R}^{n\times d}$ are continuous and there exist constants $L > 0$ and $k \geqslant 0$ such that

$$\begin{cases} |b(t,x,u) - b(t,y,u)| \leqslant L\big(1 + (|x| \vee |y|)^k + |u|\big)|x-y|, \\ \langle\, x-y, b(t,x,u) - b(t,y,u)\,\rangle \leqslant L|x-y|^2, \\ |\sigma(t,x,u) - \sigma(t,y,u)| \leqslant L|x-y|, \qquad \forall (t,u) \in [0,T]\times U,\ x,y \in \mathbb{R}^n, \end{cases}$$

with $|x| \vee |y| = \max\{|x|, |y|\}$, and

$$|b(t,0,u)| + |\sigma(t,0,u)| \leqslant L(1+|u|), \qquad \forall (t,u) \in [0,T]\times U.$$

(H2) Maps $g : D[0,T] \times \mathbb{R}^n \times U \to \mathbb{R}$ and $h : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ are continuous, and there exist constants $L > 0$ and $q \geqslant 0$ such that

$$\begin{cases} 0 \leqslant g(\tau,t,x,u) \leqslant L(1 + |x|^q + |u|^q), \\ 0 \leqslant h(\tau,x) \leqslant L(1+|x|^q), \end{cases} \qquad \forall (\tau,t,x,u) \in D[0,T]\times\mathbb{R}^n\times U.$$

It is standard that under (H1)–(H2), for any $(t,x) \in [0,T)\times\mathbb{R}^n$ and $u(\cdot) \in \mathcal{U}^{q\vee 2}[t,T]$, the state equation admits a unique solution $X(\cdot)$, and $J(t,x;u(\cdot))$ is finite for any $u(\cdot) \in \mathcal{U}^{q\vee 2}[t,T]$. We now formally state our optimal control problem.

Problem (N). For any given initial pair $(t,x) \in [0,T)\times\mathbb{R}^n$, find a $\bar{u}(\cdot) \in \mathcal{U}[t,T]$ such that

$$J(t,x;\bar{u}(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}[t,T]} J(t,x;u(\cdot)). \tag{2.2}$$

The above Problem (N) is time-inconsistent, in general. Our goal is to find *time-consistent equilibrium controls* and characterize the *equilibrium value function*, which will be made precise below.

We denote

$$a(t,x,u) = \frac{1}{2}\sigma(t,x,u)\sigma(t,x,u)^T, \qquad \forall (t,x,u) \in [0,T]\times\mathbb{R}^n\times U.$$

Define

$$\begin{aligned} \mathbb{H}(\tau,t,x,u,p,P) = \langle\, b(t,x,u), p\,\rangle + \mathrm{tr}\big[a(t,x,u)P\big] + g(\tau,t,x,u), \\ \forall (\tau,t,x,u,p,P) \in D[0,T]\times\mathbb{R}^n\times U\times\mathbb{R}^n\times\mathbb{S}^n, \end{aligned} \tag{2.3}$$

and let

$$\begin{aligned} H(\tau,t,x,p,P) = \inf_{u\in U} \mathbb{H}(\tau,t,x,u,p,P), \\ \forall (\tau,t,x,p,P) \in D[0,T]\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{S}^n. \end{aligned} \tag{2.4}$$

We introduce the following assumption.

(H3) The map $\psi : D[0,T]\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{S}^n \to U$ is well-defined and has needed regularity.

2.2. Time-consistent equilibria. In this section, inspired by [29], we will seek time-consistent solution to Problem(N) by an approach of multi-person differential games.

To begin with, let us first introduce some necessary notions. Let $\mathcal{P}[0,T]$ be the set of all partitions $\Delta = \{t_k \bigm| 0 \leqslant k \leqslant N\}$ of $[0,T]$ with $0 = t_0 < t_1 < t_2 < \cdots < t_{N-1} < t_N = T$, and with the mesh size $\|\Delta\|$ defined by the following:

$$\|\Delta\| = \max_{1\leqslant k\leqslant N}(t_k - t_{k-1}).$$

We introduce the following definition.

Definition 2.2. A map $\Psi : [0,T]\times\mathbb{R}^n \to U$ is called a *closed-loop equilibrium strategy* of Problem (N) if for any $(t,x)\in\mathcal{D}^p$, the closed-loop system

$$\begin{cases} dX^*(s) = b(s, X^*(s), \Psi(s, X^*(s)))ds + \sigma(s, X^*(s), \Psi(s, X^*(s)))dW(s), \quad s\in[t,T], \\ X^*(t) = x. \end{cases}$$

admits a unique solution $X^*(\cdot) \equiv X^*(\cdot\,; t, x, \Psi(\cdot))$. There exists a family $\mathcal{P}_0[0,T] \subseteq \mathcal{P}[0,T]$ with the property that

$$\inf_{\Delta\in\mathcal{P}_0[0,T]} \|\Delta\| = 0,$$

and for any $\Delta \equiv \{0 = t_0 < t_1 < \cdots < t_N = T\} \in \mathcal{P}_0[0,T]$, there exists a map $\Psi^\Delta : [0,T]\times\mathbb{R}^n \to U$ satisfying the following: For any $(t,x)\in\mathcal{D}^p$, the following system

$$\begin{cases} dX^\Delta(s) = b\big(s, X^\Delta(s), \Psi^\Delta(s, X^\Delta(s))\big)ds \\ \qquad\qquad +\sigma\big(s, X^\Delta(s), \Psi^\Delta(s, X^\Delta(s))\big)dW(s), \quad s\in[t,T], \\ X^\Delta(t) = x \end{cases}$$

admits a unique solution $X^\Delta(\cdot) \equiv X^\Delta(\cdot\,; t, x, \Psi^\Delta(\cdot))$. Let $X_0^\Delta(\cdot) = X^\Delta(\cdot\,; 0, x, \Psi^\Delta(\cdot))$ which is defined on $[0,T]$. For each $k = 1, 2, \cdots, N$, and any $u_k(\cdot)\in\mathcal{U}[t_{k-1}, t_k]$, let $X_k(\cdot)$ be the solution to the following:

$$\begin{cases} dX_k(s) = b(s, X_k(s), u_k(s))ds + \sigma(s, X_k(s), u_k(s))dW(s), \quad s\in[t_{k-1}, t_k], \\ dX_k(s) = b\big(s, X_k(s), \Psi^\Delta(s, X_k(s))\big)ds \\ \qquad\qquad +\sigma\big(s, X_k(s), \Psi^\Delta(s, X_k(s))\big)dW(s), \quad s\in[t_k, T], \\ X_k(t_{k-1}) = X_0^\Delta(t_{k-1}). \end{cases} \tag{2.5}$$

Then the following local optimality condition holds:

$$J\big(t_{k-1}, X_0^\Delta(t_{k-1}); \Psi^\Delta(\cdot)\big|_{[t_{k-1},T]}\big) \leqslant J(t_{k-1}, X_0^\Delta(t_{k-1}); u_k(\cdot)\oplus\Psi^\Delta(\cdot)\big|_{[t_k,T]}\big),$$

where

$$\Big(u_k(\cdot)\oplus\Psi^\Delta(\cdot)\Big|_{[t_k,T]}\Big)(s) = \begin{cases} u_k(s), & s\in[t_{k-1}, t_k), \\ \Psi^\Delta(s, X_k(s)), & s\in[t_k, T]. \end{cases} \tag{2.6}$$

Further,

$$\lim_{\Delta\in\mathcal{P}_0[0,T],\|\Delta\|\to 0} |\Psi^\Delta(t,x) - \Psi(t,x)| = 0,$$

uniformly for (t,x) in any compact set of $[0,T]\times\mathbb{R}^n$.

In the above case, we call $X_0^*(\cdot) \equiv X^*(\cdot\,;0,x,\Psi(\cdot))$ and the corresponding $u_0^*(\cdot) \equiv \Psi\big(\cdot\,;X_0^*(\cdot)\big)$ a *time-consistent equilibrium state process* and a *time-consistent equilibrium control* for the initial state x, respectively, and call $(X_0^*(\cdot),u_0^*(\cdot))$ a *time-consistent equilibrium pair*. Further, function $V:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ is called an *equilibrium value function* of Problem (N) if for each $x\in\mathbb{R}^n$,

$$V\big(t,X_0^*(t)\big)=J\big(t,X_0^*(t);\Psi(\cdot)\big|_{[t,T]}\big),\qquad t\in[0,T]. \tag{2.7}$$

We point out that the equilibrium strategy $\Psi(\cdot\,,\cdot)$ is a map defined on $[0,T]\times\mathbb{R}^n$, which is independent of particular initial pairs.

2.3. Multi-person differential games. We now consider an N-person differential game, called Problem (G^Δ), depending on the given partition $\Delta: 0=t_0<t_1<\cdots<t_N=T$ of $[0,T]$. Throughout this section, we assume that (H1)–(H3) hold. Let us start with Player N who controls the system on $[t_{N-1}t_N)$. More precisely, for each $(t,x)\in[t_{N-1},t_N]\times\mathbb{R}^n$, consider the following controlled SDE:

$$\begin{cases} dX^N(s)=b\big(s,X^N(s),u^N(s)\big)ds+\sigma\big(s,X^N(s),u^N(s)\big)dW(s),\ s\in[t,t_N],\\ X^N(t)=x, \end{cases} \tag{2.8}$$

with cost functional

$$J^N(t,x;u^N(\cdot))=\mathbb{E}_t\Big[\int_t^{t_N} g\big(t_{N-1},s,X^N(s),u^N(s)\big)ds+h\big(t_{N-1},X^N(t_N)\big)\Big]. \tag{2.9}$$

Note that

$$J^N(t_{N-1},x;u^N(\cdot))=J(t_{N-1},x;u^N(\cdot)),\quad (x,u^N(\cdot))\in\mathbb{R}^n\times\mathcal{U}[t_{N-1},t_N]. \tag{2.10}$$

We pose the following problem for Player N:

Problem ($\mathbf{C}_N$). For any $(t,x)\in[t_{N-1},t_N)\times\mathbb{R}^n$, find a $\bar u^N(\cdot)\equiv\bar u^N(\cdot\,;t,x)\in\mathcal{U}[t,t_N]$ such that

$$J^N(t,x;\bar u^N(\cdot))=\inf_{u^N(\cdot)\in\mathcal{U}[t,t_N]} J^N(t,x;u^N(\cdot))\equiv V^\Delta(t,x). \tag{2.11}$$

The above defines the *value function* $V^\Delta(\cdot\,,\cdot)$ on $[t_{N-1},t_N]\times\mathbb{R}^n$, and in the case $\bar u^N(\cdot)$ exists, by (2.10), we have

$$J(t_{N-1},x;\bar u^N(\cdot))=V^\Delta(t_{N-1},x),\qquad \forall x\in\mathbb{R}^n. \tag{2.12}$$

Under proper conditions, $V^\Delta(\cdot\,,\cdot)$ is the classical solution to the following HJB equation ([12, 18]):

$$\begin{cases} V_t^\Delta(t,x)+\inf\limits_{u\in U}\mathbb{H}\big(t_{N-1},t,x,u,V_x^\Delta(t,x),V_{xx}^\Delta(t,x)\big)=0,\\ \qquad\qquad\qquad\qquad\qquad\qquad (t,x)\in[t_{N-1},t_N]\times\mathbb{R}^n,\\ V^\Delta(t_N,x)=h(t_{N-1},x),\qquad x\in\mathbb{R}^n. \end{cases} \tag{2.13}$$

By the definition of $\psi: D[0,T]\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{S}^n\to U$, we may also write (2.13) as follows

$$\begin{cases} V_t^\Delta(t,x)+\mathbb{H}\big(t_{N-1},t,x,\psi(t_{N-1},t,x,V_x^\Delta(t,x),V_{xx}^\Delta(t,x)),V_x^\Delta(t,x),V_{xx}^\Delta(t,x)\big)=0,\\ \qquad\qquad\qquad\qquad\qquad\qquad (t,x)\in[t_{N-1},t_N]\times\mathbb{R}^n,\\ V^\Delta(t_N,x)=h(t_{N-1},x),\qquad x\in\mathbb{R}^n. \end{cases} \tag{2.14}$$

Clearly, $V^\Delta(\cdot\,,\cdot)$, well-defined on $[t_{N-1},t_N]\times\mathbb{R}^n$, depends on t_{N-1} and t_N. With such a solution $V^\Delta(\cdot\,,\cdot)$ of (2.13) (or (2.14)), let us assume that the following closed-loop system admits a unique solution $\bar X^N(\cdot)\equiv\bar X^N(\cdot\,;t_{N-1},x)$: (suppressing the dependence of $\bar X^N(\cdot)$ on t_N through $V^\Delta(\cdot\,,\cdot)$)

$$\begin{cases} d\bar X^N(s)=b\big(s,\bar X^N(s),\psi(t_{N-1},s,\bar X^N(s),V_x^\Delta(s,\bar X^N(s)),V_{xx}^\Delta(s,\bar X^N(s))\big)ds\\ \qquad\qquad +\sigma\big(s,\bar X^N(s),\psi(t_{N-1},s,\bar X^N(s),V_x^\Delta(s,\bar X^N(s)),V_{xx}^\Delta(s,\bar X^N(s))\big)dW(s),\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s\in[t_{N-1},t_N],\\ \bar X^N(t_{N-1})=x. \end{cases} \tag{2.15}$$

Then under (H3), an optimal control $\bar u^N(\cdot)$ of Problem (C_N) for the initial pair (t_{N-1},x) admits the following feedback representation:

$$\begin{aligned} \bar u^N(s)&\equiv\bar u^N(s;t_{N-1},x)=\psi\big(t_{N-1},s,\bar X^N(s),V_x^\Delta(s,\bar X^N(s)),V_{xx}^\Delta(s,\bar X^N(s))\big)\\ &\equiv\psi\big(t_{N-1},s,\bar X^N(s;t_{N-1},x),V_x^\Delta(s,\bar X^N(s;t_{N-1},x)),V_{xx}^\Delta(s,\bar X^N(s;t_{N-1},x))\big)\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s\in[t_{N-1},t_N], \end{aligned} \tag{2.16}$$

and $\bar X^N(\cdot)\equiv\bar X^N(\cdot\,;t_{N-1},x)$ is the corresponding optimal state process.

Next, we consider an optimal control problem for Player $(N-1)$ on $[t_{N-2},t_{N-1})$. For any initial pair $(t,x)\in[t_{N-2},t_{N-1}]\times\mathbb{R}^n$, the state equation is

$$\begin{cases} dX^{N-1}(s)=b\big(s,X^{N-1}(s),u^{N-1}(s)\big)ds+\sigma\big(s,X^{N-1}(s),u^{N-1}(s)\big)dW(s),\\ \qquad\qquad\qquad\qquad\qquad\qquad s\in[t,t_{N-1}),\\ X^{N-1}(t)=x. \end{cases} \tag{2.17}$$

To determine the suitable cost functional, we note that Player $(N-1)$ can only control the system on $[t_{N-2},t_{N-1})$ and Player N will take over at t_{N-1} to control the system thereafter. Moreover, Player $(N-1)$ knows that Player N will play optimally based on the initial pair $(t_{N-1},X^{N-1}(t_{N-1}))$ of Player N, which is the *terminal pair* of Player $(N-1)$. Hence, the *sophisticated cost functional* of Player $(N-1)$ should be

$$\begin{aligned} J^{N-1}(t,x;u^{N-1}(\cdot))&=\mathbb{E}_t\Big[\int_t^{t_{N-1}}g(t_{N-2},s,X^{N-1}(s),u^{N-1}(s))ds\\ &+\int_{t_{N-1}}^{t_N}g\big(t_{N-2},s,\bar X^N(s;t_{N-1},X^{N-1}(t_{N-1})),\bar u^N(s;t_{N-1},X^{N-1}(t_{N-1}))\big)ds\\ &+h\big(t_{N-2},\bar X^N(t_N;t_{N-1},X^{N-1}(t_{N-1}))\big)\Big]. \end{aligned} \tag{2.18}$$

Note that although Player $(N-1)$ knows that Player N will control the system on $[t_{N-1},t_N]$, he/she still "discounts" the future costs in his/her own way (see t_{N-2} appearing in the running cost on $[t_{N-1},t_N]$ and in the terminal cost at t_N). Now, if we denote

$$\begin{aligned} h^{N-1}(x)=\mathbb{E}_{t_{N-1}}\Big[&\int_{t_{N-1}}^{t_N}g\big(t_{N-2},s,\bar X^N(s;t_{N-1},x),\bar u^N(s;t_{N-1},x)\big)ds\\ &+h\big(t_{N-2},\bar X^N(t_N;t_{N-1},x)\big)\Big], \end{aligned}$$

then the cost functional (2.18) can be written as

$$\begin{aligned} J^{N-1}(t,x;u^{N-1}(\cdot))=\mathbb{E}_t\Big[&\int_t^{t_{N-1}}g\big(t_{N-2},s,X^{N-1}(s),u^{N-1}(s)\big)ds\\ &+h^{N-1}(X^{N-1}(t_{N-1}))\Big]. \end{aligned} \tag{2.19}$$

We see that the optimal control problem associated with the state equation (2.17) and the cost functional (2.19) looks like a standard one. But, the map $x \mapsto h^{N-1}(x)$ seems to be a little too implicit, which is difficult for us to pass to the limits later on. We now would like to make it more explicit in some sense. Inspired by the idea of Four Step Scheme introduced in [21, 22] for forward-backward stochastic differential equations (FBSDEs, for short) with deterministic coefficients, we proceed as follows. For the optimal state process $\bar{X}^N(\cdot) \equiv \bar{X}^N(\cdot\,; t_{N-1}, x)$ determined by (2.15) on $[t_{N-1}, t_N]$, we introduce the following backward stochastic differential equation (BSDE, for short):

$$\begin{cases} dY^N(s) = -g\big(t_{N-2}, s, \bar{X}^N(s), \psi(t_{N-1}, s, \bar{X}^N(s), V_x^\Delta(s, \bar{X}^N(s)), V_{xx}^\Delta(s, \bar{X}^N(s)))\big)ds \\ \qquad\qquad + Z^N(s)dW(s), \qquad\qquad s \in [t_{N-1}, t_N], \\ Y^N(t_N) = h(t_{N-2}, \bar{X}^N(t_N)), \end{cases} \tag{2.20}$$

which is equivalent to the following:

$$\begin{cases} dY^N(s) = -g\big(t_{N-2}, s, \bar{X}^N(s), \bar{u}^N(s)\big)ds + Z^N(s)dW(s), \ s \in [t_{N-1}, t_N], \\ Y^N(t_N) = h(t_{N-2}, \bar{X}^N(t_N)), \end{cases} \tag{2.21}$$

Note that t_{N-2} appears in the drift of BSDE and in the terminal condition. This BSDE admits a unique adapted solution $(Y^N(\cdot), Z^N(\cdot)) \equiv \big(Y^N(\cdot\,; x), Z^N(\cdot\,; x)\big)$ ([22, 39]), uniquely depending on $x \in \mathbb{R}^n$. Further, one has

$$\begin{aligned} Y^N(t_{N-1}) &= \mathbb{E}_{t_{N-1}}\Big[\int_{t_{N-1}}^{t_N} g\big(t_{N-2}, s, \bar{X}^N(s), \bar{u}^N(s)\big)ds + h(t_{N-2}, \bar{X}^N(t_N))\Big] \\ &= h^{N-1}(x). \end{aligned}$$

It is seen that (2.15) and (2.21) form an FBSDE. By [21] (see also [22, 39]), we have the following representation for $Y^N(\cdot)$

$$Y^N(s) = \Theta^N(s, \bar{X}^N(s)), \qquad s \in [t_{N-1}, t_N], \tag{2.22}$$

as long as $\Theta^N(\cdot\,,\cdot)$ is a classical solution to the following PDE:

$$\begin{cases} \Theta_s^N(s,x) + \langle\, \Theta_x^N(s,x), b\big(s, x, \psi(t_{N-1}, s, x, V_x^\Delta(s,x), V_{xx}^\Delta(s,x))\big)\,\rangle \\ \qquad + \mathrm{tr}\,\big[a\big(s, x, \psi(t_{N-1}, s, x, V_x^\Delta(s,x), V_{xx}^N(s,x))\big)\Theta_{xx}^N(s,x)\big] \\ \qquad + g\big(t_{N-2}, s, x, \psi(t_{N-1}, s, x, V_x^\Delta(s,x), V_{xx}^\Delta(s,x))\big) = 0, \\ \qquad\qquad\qquad\qquad (s,x) \in [t_{N-1}, t_N] \times \mathbb{R}^n, \\ \Theta^N(t_N, x) = h(t_{N-2}, x). \qquad x \in \mathbb{R}^n, \end{cases} \tag{2.23}$$

Note that $\Theta^N(\cdot\,,\cdot)$ depends on (t_{N-2}, t_{N-1}, t_N). With the above representation $\Theta^N(\cdot\,,\cdot)$ of $Y^N(\cdot)$, we can rewrite the cost functional (2.19) as follows:

$$\begin{aligned} &J^{N-1}(t, x; u^{N-1}(\cdot)) \\ &= \mathbb{E}_t\Big[\int_t^{t_{N-1}} g(t_{N-2}, s, X^{N-1}(s), u^{N-1}(s))ds + \Theta^N(t_{N-1}, X^{N-1}(t_{N-1}))\Big]. \end{aligned}$$

We now pose the following problem for Player $(N-1)$:

Problem ($\mathbf{C}_{N-1}$). For any $(t,x) \in [t_{N-2}, t_{N-1}) \times \mathbb{R}^n$, find a $\bar{u}^{N-1}(\cdot) \equiv \bar{u}^{N-1}(\cdot\,; t, x) \in \mathcal{U}[t_{N-2}, t_{N-1}]$ such that

$$J^{N-1}(t, x; \bar{u}^{N-1}(\cdot)) = \inf_{u^{N-1}(\cdot) \in \mathcal{U}[t, t_{N-1}]} J^{N-1}(t, x; u^{N-1}(\cdot)) \equiv V^\Delta(t, x). \tag{2.24}$$

The above defines the value function $V^\Delta(\cdot\,,\cdot)$ on $[t_{N-2},t_{N-1})\times\mathbb{R}^n$. Under proper conditions, $V^\Delta(\cdot\,,\cdot)$ is the classical solution to the following HJB equation ([12, 18]):

$$\begin{cases} V_t^\Delta(t,x)+\inf\limits_{u\in U}\mathbb{H}\big(t_{N-2},t,x,u,V_x^\Delta(t,x),V_{xx}^\Pi(t,x)\big)=0, \\ \qquad\qquad\qquad\qquad\qquad\qquad (t,x)\in[t_{N-2},t_{N-1})\times\mathbb{R}^n, \\ V^\Delta(t_{N-1}-0,x)=\Theta^N(t_{N-1},x), \qquad x\in\mathbb{R}^n. \end{cases} \tag{2.25}$$

We point out that in general,

$$V^\Delta(t_{N-1}-0,x)=\Theta^N(t_{N-1},x)\neq V^\Delta(t_{N-1},x).$$

Thus, $V^\Delta(\cdot\,,\cdot)$, which is now defined on $[t_{N-2},t_N]\times\mathbb{R}^n$, may have a jump at $\{t_{N-1}\}\times\mathbb{R}^n$. For any $x\in\mathbb{R}^n$, suppose the following admits a unique solution $\bar X^{N-1}(\cdot)$:

$$\begin{cases} d\bar X^{N-1}(s)=b\big(s,\bar X^{N-1}(s),\psi^{N-1}(s)\big)ds+\sigma\big(s,\bar X^{N-1}(s),\psi^{N-1}(s)\big)dW(s), \\ \psi^{N-1}(s)=\psi\big(t_{N-2},s,\bar X^{N-1}(s),V_x^\Delta(s,\bar X^{N-1}(s)),V_{xx}^\Delta(s,\bar X^{N-1}(s))\big), \\ \qquad\qquad\qquad\qquad\qquad\qquad s\in[t_{N-2},t_{N-1}), \\ \bar X^{N-1}(t_{N-2})=x. \end{cases} \tag{2.26}$$

Then we define

$$\bar u^{N-1}(s;t_{N-2},x)=\psi\big(t_{N-2},s,\bar X^{N-1}(s),V_x^\Delta(s,\bar X^{N-1}(s)),V_{xx}^\Delta(s,\bar X^{N-1}(s))\big), \quad s\in[t_{N-2},t_{N-1}),$$

which is an optimal control of Problem (C_{N-1}) with the initial pair (t_{N-2},x). Now, for the optimal pair

$$\big(\bar X^{N-1}(\cdot),\bar u^{N-1}(\cdot)\big)=\big(\bar X^{N-1}(\cdot\,;t_{N-2},x),\bar u^{N-1}(\cdot\,;t_{N-2},x)\big)$$

of Problem (C_{N-1}) (on $[t_{N-2},t_{N-1}]$), we make a natural extension on $[t_{N-1},t_N]$ as follows:

$$\begin{cases} \bar X^{N-1}(s)=\bar X^N(s;t_{N-1},\bar X^{N-1}(t_{N-1})), \\ \bar u^{N-1}(s)=\bar u^N(s;t_{N-1},\bar X^{N-1}(t_{N-1})), \end{cases} \qquad s\in[t_{N-1},t_N].$$

We refer to such a pair $\big(\bar X^{N-1}(\cdot),\bar u^{N-1}(\cdot)\big)$ as a *sophisticated equilibrium pair* on $[t_{N-2},t_N]$.

The above procedure can be continued recursively. By induction, we can construct sophisticated cost functional $J^k(t,x;u^k(\cdot))$ for Player k, and

$$V^\Delta(t,x)=\inf_{u^k(\cdot)\in\mathcal{U}[t,t_k]}J^k(t,x;u^k(\cdot)), \qquad (t,x)\in[t_{k-1},t_k)\times\mathbb{R}^n,\ 1\leqslant k\leqslant N,$$

with the value function $V^\Delta(\cdot\,,\cdot)$ satisfying the following HJB equations on the time intervals associated with the partition Δ:

$$\begin{cases} V_t^\Delta(t,x)+\mathbb{H}\big(\ell^\Delta(t),t,x,\psi\big(\ell^\Delta(t),t,x,V_x^\Delta(t,x),V_{xx}^\Delta(t,x)\big),V_x^\Delta(t,x),V_{xx}^\Delta(t,x)\big)=0, \\ \qquad\qquad\qquad\qquad\qquad\qquad (t,x)\in[t_{N-1},t_N]\times\mathbb{R}^n, \\ V^\Delta(t_N,x)=h(t_{N-1},x), \qquad x\in\mathbb{R}^n, \end{cases}$$

and for $k=1,2,\cdots,N-1$,

$$\begin{cases} V_t^\Delta(t,x)+\mathbb{H}\big(\ell^\Delta(t),t,x,\psi\big(\ell^\Delta(t),t,x,V_x^\Delta(t,x),V_{xx}^\Delta(t,x)\big),V_x^\Delta(t,x),V_{xx}^\Delta(t,x)\big)=0, \\ \qquad\qquad\qquad\qquad\qquad\qquad (t,x)\in[t_{k-1},t_k)\times\mathbb{R}^n, \\ V^\Delta(t_k-0,x)=\Theta^{k+1}(t_k,x), \qquad x\in\mathbb{R}^n, \end{cases}$$

where,

$$\ell^{\Delta}(s)=\sum_{k=1}^{N}t_{k-1}I_{[t_{k-1},t_k)}(s),\qquad s\in[0,T],$$

and for $k=1,2,\cdots,N-1$, $\Theta^{k+1}(\cdot\,,\cdot)$ is the solution to the following (linear) PDE:

$$\begin{cases}\Theta_t^{k+1}(t,x)+\mathbb{H}\big(t_{k-1},t,x,\psi\big(\ell^{\Delta}(t),t,x,V_x^{\Delta}(t,x),V_{xx}^{\Delta}(t,x)\big),\\ \qquad\qquad\qquad\Theta_x^{k+1}(t,x),\Theta_{xx}^{k+1}(t,x)\big)=0,\quad (t,x)\in[t_k,t_N]\times\mathbb{R}^n,\\ \Theta^{k+1}(t_N,x)=h(t_{k-1},x),\qquad\qquad x\in\mathbb{R}^n.\end{cases}\tag{2.27}$$

Now, we define

$$\Psi^{\Delta}(t,x)=\psi\big(\ell^{\Delta}(t),t,x,V_x^{\Delta}(t,x),V_{xx}^{\Delta}(t,x)\big),\qquad (t,x)\in[0,T]\times\mathbb{R}^n.\tag{2.28}$$

Then for any given $x\in\mathbb{R}^n$, let $X_0^{\Delta}(\cdot)$ be the solution to the following closed-loop system:

$$\begin{cases}dX_0^{\Delta}(s)=b\big(s,X_0^{\Delta}(s),\Psi^{\Delta}(s,X_0^{\Delta}(s))\big)ds+\sigma\big(s,X_0^{\Delta}(s),\Psi^{\Delta}(s,\bar{X}_0^{\Delta}(s))\big)dW(s),\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s\in[0,T],\\ X_0^{\Delta}(0)=x,\end{cases}$$

and denote

$$u_0^{\Delta}(s)=\Psi^{\Delta}\big(s,X_0^{\Delta}(s)\big),\qquad s\in[0,T].$$

According to our construction, we have

$$\begin{aligned}&J\big(t_{k-1},X_0^{\Delta}(t_{k-1});\Psi^{\Delta}(\cdot)\big|_{[t_{k-1},T]}\big)=J\big(t_{k-1},X_0^{\Delta}(t_{k-1});u_0^{\Delta}(\cdot)\big|_{[t_{k-1},t_k]}\big)\\ &=V^{\Delta}\big(t_{k-1},X_0^{\Delta}(t_{k-1})\big)=J^k\big(t_{k-1},X_0^{\Delta}(t_{k-1});u_0^{\Delta}(\cdot)\big|_{[t_{k-1},t_k]}\big)\\ &=\inf_{u^k(\cdot)\in\mathcal{U}[t_{k-1},t_k]}J^k\big(t_{k-1},X_0^{\Delta}(t_{k-1});u^k(\cdot)\big)\leqslant J^k\big(t_{k-1},X_0^{\Delta}(t_{k-1});u^k(\cdot)\big)\\ &=J\big(t_{k-1},X_0^{\Delta}(t_{k-1});u^k(\cdot)\oplus\Psi^{\Delta}(\cdot)\big|_{[t_k,T]}\big),\\ &\qquad\qquad\qquad\qquad\forall u^k(\cdot)\in\mathcal{U}[t_{k-1},t_k],\quad 1\leqslant k\leqslant N,\end{aligned}\tag{2.29}$$

where $u^k(\cdot)\oplus\Psi^{\Delta}(\cdot)\big|_{[t_k,T]}$ is defined the same way as (2.5)–(2.6). In general, for $k=1,2,\cdots,N-1$, we might have

$$J\big(t_{k-1},X_0^{\Delta}(t_{k-1});u_0^{\Delta}(\cdot)\big)>\inf_{u(\cdot)\in\mathcal{U}[t_{k-1},T]}J(t_{k-1},X_0^{\Delta}(t_{k-1});u(\cdot)\big).\tag{2.30}$$

2.4. The limits. We now would like to look at the situation when $\|\Delta\|\to 0$. Suppose we have the following:

$$\lim_{\|\Delta\|\to 0}\Big(|V^{\Delta}(t,x)-V(t,x)|+|V_x^{\Delta}(t,x)-V_x(t,x)|+|V_{xx}^{\Delta}(t,x)-V_{xx}(t,x)|\Big)=0,$$

uniformly for (t,x) in any compact sets, for some $V(\cdot\,,\cdot)$. Under (H3), we also have

$$\lim_{\|\Delta\|\to 0}|\Psi^{\Delta}(t,x)-\Psi(t,x)|=0,$$

uniformly for (t,x) in any compact sets, for

$$\Psi(t,x)=\psi\big(t,t,x,V_x(t,x),V_{xx}(t,x)\big),\qquad (t,x)\in[0,T]\times\mathbb{R}^n.\tag{2.31}$$

Then the following limit exist:

$$\lim_{\|\Delta\|\to 0}\|X_0^\Delta(\cdot)-X^*(\cdot)\|_{L^2_{\mathbb{F}}(\Omega;C([0,T];\mathbb{R}^n))}=0,$$

for $X^*(\cdot)$ solving the following SDE:

$$\begin{cases} dX^*(s)=b\big(s,X^*(s),u^*(s)\big)ds+\sigma\big(s,X^*(s),u^*(s)\big)dW(s), \quad s\in[0,T].\\ X^*(0)=x, \end{cases} \tag{2.32}$$

where

$$u^*(s)=\Psi\big(s,X^*(s)\big), \qquad s\in[0,T], \tag{2.33}$$

and

$$L^2_{\mathbb{F}}(\Omega,C([0,T];\mathbb{R}^n))=\Big\{X:[0,T]\times\Omega\to\mathbb{R}^n\ \Big|\ X(\cdot)\text{ has continuous paths},\ \mathbb{E}\big[\sup_{t\in[0,T]}|X(t)|^2\big]<\infty\Big\}.$$

Clearly,

$$\lim_{\|\Delta\|\to 0}\|u_0^\Delta(\cdot)-u^*(\cdot)\|_{\mathcal{U}^2[0,T]}=0.$$

By (2.29), we have

$$J\big(\ell^\Delta(t),\bar X^\Delta(\ell^\Delta(t));\bar u^\Delta(\cdot)\big)=V^\Delta(\ell^\Delta(t),\bar X^\Delta(t)), \qquad t\in[0,T].$$

Thus, passing to the limits, we have (2.7).

By Definition 2.2, $\Psi(\cdot\,,\cdot)$ is a time-consistent equilibrium strategy, and $V(\cdot\,,\cdot)$ is a time-consistent equilibrium value function of Problem (N).

With some careful analysis (see [36]), we are able to obtain the following differential equation:

$$\begin{cases} \Theta_t(\tau,t,x)+\mathbb{H}\big(\tau,t,x,\psi\big(t,t,x,\Theta_x(t,t,x),\Theta_{xx}(t,t,x)\big),\Theta_x(\tau,t,x),\Theta_{xx}(\tau,t,x)\big)=0,\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\tau,t,x)\in D[0,T]\times\mathbb{R}^n,\\ \Theta(\tau,T,x)=h(\tau,x), \qquad (\tau,x)\in[0,T]\times\mathbb{R}^n, \end{cases} \tag{2.34}$$

where

$$\begin{cases} \mathbb{H}(\tau,t,x,u,p,P)=\text{tr}\,\big[a(t,x,u)P\big]+\langle\, b(t,x,u),p\,\rangle+g(\tau,t,x,u),\\ \psi(\tau,t,x,p,P)\in\arg\min\mathbb{H}(\tau,t,x,\cdot\,,p,P). \end{cases} \tag{2.35}$$

We call the above (2.34) the *equilibrium Hamilton-Jacobi-Bellman equation* (equilibrium HJB equation, for short) of Problem (N). If one can find $\Theta(\cdot\,,\cdot\,,\cdot)$ from the above, then the equilibrium value function $V(\cdot\,,\cdot)$ can be determined by the following:

$$V(t,x)=\Theta(t,t,x), \qquad (t,x)\in[0,T]\times\mathbb{R}^n. \tag{2.36}$$

It is clear that the (time-consistent) equilibrium pair can be determined by (2.32) and (2.33), in principle.

An interesting feature of (2.34) that both $\Theta(\tau,t,x)$ and $\Theta(t,t,x)$ appear in the equation where the later is the restriction of the former on $\tau=t$. On one hand, although the equation is fully nonlinear, due to the fact that $\Theta(t,t,x)$ is different from $\Theta(\tau,t,x)$, the existing theory for fully nonlinear parabolic equations cannot apply directly. On the other hand, it is seen that if $\Theta(t,t,x)$ is obtained from an independent way, then (2.34) is actually a linear equation for $\Theta(\tau,t,x)$ with τ can be purely regarded as a parameter.

2.5. Well-posedness of the equilibrium HJB equation. In this subsection, we discuss the well-posedness for the equilibrium HJB equation (2.34). Let us first intuitively describe our idea. For any smooth function $v(\cdot\,,\cdot)$, denote

$$\begin{cases} \big[\mathcal{L}\big(t,v(t,\cdot)\big)\varphi(\cdot)\big](x) = \text{tr}\,\big[a(t,x,\psi(t,t,x,v_x(t,x),v_{xx}(t,x))\varphi_{xx}(x)\big] \\ \qquad\qquad + \langle\, b\big(t,x,\psi(t,t,x,v_x(t,x),v_{xx}(t,x))\big),\varphi_x(x)\,\rangle, \\ \qquad\qquad\qquad (t,x)\in[0,T]\times\mathbb{R}^n, \\ \mathcal{G}(\tau,t,v(t,\cdot))(x) = g\big(\tau,t,x,\psi(t,t,x,v_x(t,x),v_{xx}(t,x))\big), \quad (\tau,t,x)\in D[0,T]\times\mathbb{R}^n. \end{cases}$$

Consider the following linear abstract backward evolution equation:

$$\begin{cases} \Theta_t(\tau,t) + \mathcal{L}(t,v(t))\Theta(\tau,t) + \mathcal{G}(\tau,t,v(t)) = 0, \qquad t\in[\tau,T], \\ \Theta(\tau,T) = h(\tau). \end{cases} \tag{2.37}$$

Under some mild conditions, the above is well-posed, and we have the following variation of constant formula:

$$\Theta(\tau,t) = \mathcal{E}(T,t;v(\cdot))h(\tau) + \int_t^T \mathcal{E}(s,t;v(\cdot))\mathcal{G}(\tau,s,v(s))ds, \qquad t\in[\tau,T], \tag{2.38}$$

where $\mathcal{E}(\cdot\,,\cdot\,;v(\cdot))$ is called the *backward evolution operator* generated by $\mathcal{L}(\cdot\,,v(\cdot))$. Consequently, the (time-consistent) equilibrium value function $V(t,\cdot)=\Theta(t,t,\cdot)$ should be the solution to the following nonlinear functional integral equation:

$$V(t) = \mathcal{E}(T,t;V(\cdot))h(t) + \int_t^T \mathcal{E}(s,t;V(\cdot))\mathcal{G}(t,s,V(s))ds, \qquad t\in[0,T]. \tag{2.39}$$

We call (2.39) the *equilibrium HJB integral equation* for Problem (N). Once a solution $V(\cdot\,,\cdot)$ of (2.39) is found, we can, in principle, construct a (time-consistent) equilibrium control and an equilibrium pair for Problem (N). Of course, if we like, we may also solve the equilibrium HJB equation (2.34), which actually is not necessary as far as the construction of a time-consistent equilibrium pair is concerned.

The well-posedness of (2.39) seems to be difficult for the general case. We now assume the following:

$$\sigma(t,x,u) = \sigma(t,x), \qquad (t,x,u)\in[0,T]\times\mathbb{R}^n\times U, \tag{2.40}$$

namely, the control does not enter the diffusion of the state equation. In this case, our equilibrium HJB equation reads

$$\begin{cases} \Theta_t(\tau,t,x) + \dfrac{1}{2}\text{tr}\,\big[\sigma(t,x)\sigma(t,x)^T\Theta_{xx}(\tau,t,x)\big] \\ \qquad + \langle\, b\big(t,x,\psi(t,t,x,\Theta_x(t,t,x))\big),\Theta_x(\tau,t,x)\,\rangle \\ \qquad + g\big(\tau,t,x,\psi(t,t,x,\Theta_x(t,t,x))\big) = 0, \quad (\tau,t,x)\in D[0,T]\times\mathbb{R}^n, \\ \Theta(\tau,T,x) = h(\tau,x), \qquad (\tau,x)\in[0,T]\times\mathbb{R}^n. \end{cases} \tag{2.41}$$

The essential feature of (2.41) is that $\Theta_{xx}(t,t,x)$ does not appear in the equation (although $\Theta_x(t,t,x)$ still appears there). This leads to the well-posedness problem much more accessible. We have the following result (see [36]).

Theorem 2.5. *Let all the coefficients of* (2.41) *have all required order differentiability with bounded derivatives. Let*

$$a(t,x) \geqslant \delta I, \qquad (t,x)\in[0,T]\times\mathbb{R}^n,$$

for some $\delta>0$. *Then* (2.41) *admits a unique solution* $\Theta(\cdot\,,\cdot\,,\cdot)$.

3. LQ problem with nonlinearly appearance of conditional expectations

We now consider the case that conditional expectation of the state/control nonlinearly appear in the cost functional. For such a case, we will only consider linear-quadratic case. The general nonlinear situation is still open. We consider the following controlled linear SDE:

$$\begin{cases} dX(s) = \Big\{A(s)X(s) + B(s)u(s)\Big\}ds + \Big\{C(s)X(s)+D(s)u(s)\Big\}dW(s), \quad s \in [t,T], \\ X(t) = x \in \mathcal{X}_t, \end{cases} \tag{3.1}$$

Note that for any $(t,x) \in \mathcal{D}$ and $u(\cdot) \in \mathcal{U}[t,T]$, the corresponding state process $X(\cdot) = X(\cdot\,;t,x,u(\cdot))$ depends on $(t,x,u(\cdot))$. The cost functional is as follows:

$$\begin{aligned} J(t,x;u(\cdot)) = \mathbb{E}_t\Big\{ \int_t^T \big[\langle\, Q(s,t)X(s), X(s)\,\rangle + \langle\, \bar{Q}(s,t)\mathbb{E}_t[X(s)], \mathbb{E}_t[X(s)]\,\rangle \\ + \langle\, R(s,t)u(s), u(s)\,\rangle + \langle\, \bar{R}(s,t)\mathbb{E}_t[u(s)], \mathbb{E}_t[u(s)]\,\rangle\,\big]ds \\ + \langle\, G(t)X(T), X(T)\,\rangle + \langle\, \bar{G}(t)\mathbb{E}_t[X(T)], \mathbb{E}_t[X(T)]\,\rangle\Big\}. \end{aligned} \tag{3.2}$$

Let us introduce the following hypotheses:

(LQ1) The following hold:

$$A(\cdot), C(\cdot) \in C([0,T];\mathbb{R}^{n\times n}), \qquad B(\cdot), D(\cdot) \in C([0,T];\mathbb{R}^{n\times m}).$$

(LQ2) The following hold:

$$\begin{cases} Q(\cdot\,,\cdot), \bar{Q}(\cdot\,,\cdot) \in C([0,T]^2;\mathbb{S}^n), \quad R(\cdot\,,\cdot), \bar{R}(\cdot\,,\cdot) \in C([0,T]^2;\mathbb{S}^m), \\ G(\cdot), \bar{G}(\cdot) \in C([0,T];\mathbb{S}^n), \end{cases}$$

and for some $\delta > 0$,

$$\begin{cases} Q(s,t),\ Q(s,t) + \bar{Q}(s,t) \geqslant 0, \quad R(s,t),\ R(s,t) + \bar{R}(s,t) \geqslant \delta I, \quad 0 \leqslant t \leqslant s \leqslant T, \\ G(t),\ G(t) + \bar{G}(t) \geqslant 0, \qquad 0 \leqslant t \leqslant T. \end{cases}$$

(LQ3) The following monotonicity conditions are satisfied:

$$\begin{cases} Q(s,t) \leqslant Q(s,\tau), \quad Q(s,t) + \bar{Q}(s,t) \leqslant Q(s,\tau) + \bar{Q}(s,\tau), \\ R(s,t) \leqslant R(s,\tau), \quad R(s,t) + \bar{R}(s,t) \leqslant R(s,\tau) + \bar{R}(s,\tau), \qquad 0 \leqslant t \leqslant \tau \leqslant s \leqslant T. \\ G(t) \leqslant G(\tau), \quad G(t) + \bar{G}(t) \leqslant G(\tau) + \bar{G}(\tau), \end{cases}$$

It is clear that under (LQ1)–(LQ2), for any $(t,x) \in \mathcal{D}$ and $u(\cdot) \in \mathcal{U}[t,T]$, state equation (3.1) admits a unique solution $X(\cdot) \equiv X(\cdot\,;t,x,u(\cdot))$, and the cost functional $J(t,x;u(\cdot))$ is well-defined. Then we can state the following problem.

Problem (LQ). For any $(t,x) \in \mathcal{D}$, find a $u^*(\cdot) \in \mathcal{U}[t,T]$ such that

$$J(t,x;u^*(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}[t,T]} J(t,x;u(\cdot)) \equiv V(t,x). \tag{3.3}$$

For given $(t,x) \in \mathcal{D}$, any $u^*(\cdot) \in \mathcal{U}[t,T]$ satisfying the above is called a *pre-commitment optimal control* for Problem (LQ) at (t,x). The corresponding $X^*(\cdot)$ and $(X^*(\cdot), u^*(\cdot))$ are called *pre-commitment optimal state process* and *pre-commitment optimal pair* of Problem (LQ), respectively, and $V(\cdot\,,\cdot)$ is called the *pre-commitment value function*.

In what follows, we will denote

$$\begin{cases} \widehat{Q}(s,t) = Q(s,t) + \bar{Q}(s,t), \quad \widehat{R}(s,t) = R(s,t) + \bar{R}(s,t), \\ \widehat{G}(t) = G(t) + \widehat{G}(t). \end{cases} \qquad 0 \leqslant t \leqslant s \leqslant T.$$

The following is found in [37].

Proposition 3.2. *Let* (LQ1)–(LQ2) *hold. Then for any fixed $t \in [0,T)$, the following Riccati equation system admits a unique solution $(P(\cdot), \widehat{P}(\cdot)) \in C^1([t,T];\mathbb{S}^n)^2$ (suppressing s):*

$$\begin{cases} \dot{P} + PA + A^T P + C^T PC + Q(t) \\ -(PB + C^T PD)\big[R(t) + D^T PD\big]^{-1}(B^T P + D^T PC) = 0, \\ \dot{\widehat{P}} + \widehat{P}A + A^T \widehat{P} + C^T PC + \widehat{Q}(t) \\ -(\widehat{P}B + C^T PD)\big[\widehat{R}(t) + D^T PD\big]^{-1}(B^T \widehat{P} + D^T PC) = 0, \quad s \in [t,T], \\ P(T) = G(t), \quad \widehat{P}(T) = \widehat{G}(t). \end{cases} \tag{3.4}$$

Further, let $x \in \mathcal{X}_t$ and $X^*(\cdot) \equiv X^*(\cdot\,;t,x)$ be the solution to the following closed-loop system:

$$\begin{cases} dX^*(s) = \Big\{\big[A(s) - B(s)\Theta(s)\big]X^*(s) + B(s)\big[\Theta(s) - \widehat{\Theta}(s)\big]\mathbb{E}_t[X^*(s)]\Big\}ds \\ \qquad\qquad + \Big\{\big[C(s) - D(s)\Theta(s)\big]X^*(s) + D(s)\big[\Theta(s) - \widehat{\Theta}(s)\big]\mathbb{E}_t[X^*(s)]\Big\}dW(s), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s \in [t,T], \\ X^*(t) = x, \end{cases}$$

with

$$\begin{cases} \Theta(s) = \big[R(s,t) + D(s)^T P(s) D(s)\big]^{-1}\big[B(s)^T P(s) + D(s)^T P(s) C(s)\big], \\ \widehat{\Theta}(s) = \big[\widehat{R}(s,t) + D(s)^T P(s) D(s)\big]^{-1}\big[B(s)^T \widehat{P}(s) + D(s)^T P(s) C(s)\big], \end{cases}$$

and define $u^*(\cdot)$ as follows:

$$u^*(s) = -\Theta(s)X^*(s) + \big[\Theta(s) - \widehat{\Theta}(s)\big]\mathbb{E}_t[X^*(s)], \qquad s \in [t,T]. \tag{3.5}$$

Then $(X^*(\cdot), u^*(\cdot))$ is the pre-commitment optimal pair of Problem (LQ) *at (t,x), and*

$$V(t,x) = \inf_{u(\cdot)\in\mathcal{U}[t,T]} J(t,x;u(\cdot)) = J(t,x;u^*(\cdot)) = \langle\, \widehat{P}(t)x, x\,\rangle, \quad \forall x \in \mathcal{X}_t. \tag{3.6}$$

We note that the equation for $\widehat{P}(\cdot)$ can also be written

$$\begin{cases} \dot{\widehat{P}}(s) + \widehat{P}(s)A(s) + A(s)^T \widehat{P} + C(s)^T P(s) C(s) + \widehat{Q}(t,s) \\ \quad -\big[\widehat{P}(s)B(s) + C(s)^T P(s) D(s)\big]\big[\widehat{R}(s,t) + D(s)^T P(s) D(s)\big]^{-1} \\ \qquad\qquad \cdot\big[B(s)^T \widehat{P}(s) + D(s)^T P(s) C(s)\big] = 0, \qquad s \in [t,T], \\ \widehat{P}(T) = \widehat{G}(t), \end{cases} \tag{3.7}$$

Thus, as long as

$$\bar{Q}(\cdot\,,\cdot) = 0, \qquad \bar{R}(\cdot\,,\cdot) = 0, \qquad \bar{G}(\cdot) = 0$$

are not true,

$$\Theta(s) = \widehat{\Theta}(s), \qquad s \in [t,T]$$

is not true in general. Hence, the term $\mathbb{E}_t[X^*(\cdot)]$ will present in the state feedback representation of $u^*(\cdot)$ (see (3.5)), and the closed-loop system reads

$$\begin{cases} dX^*(s) = \Big\{\big[A(s)-B(s)\Theta(s)\big]X^*(s)+B(s)\big[\Theta(s)-\widehat{\Theta}(s)\big]\mathbb{E}_t[X^*(s)]\Big\}ds \\ \qquad\qquad +\Big\{\big[C(s)-D(s)\Theta(s)\big]X^*(s)+D(s)\big[\Theta(s)-\widehat{\Theta}(s)\big]\mathbb{E}_t[X^*(s)]\Big\}dW(s), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s\in[t,T], \\ X^*(t) = x, \end{cases}$$

which is a linear MF-SDE.

3.1. Open-loop equilibrium control. We introduce the following notion.

Definition 3.3. For given $x \in \mathbb{R}^n$, a state-control pair $(X^*(\cdot), u^*(\cdot)) \in \mathcal{X}[0,T] \times \mathcal{U}[0,T]$ is called an *open-loop equilibrium pair* of Problem (MF-LQ) for the initial state x if

$$X^*(0) = x,$$

and for almost all $t \in [0,T)$, and any $u(\cdot) \in \mathcal{U}[t,T]$,

$$\varliminf_{\varepsilon \downarrow 0} \frac{J(t, X^*(t); u^\varepsilon(\cdot)) - J(t, X^*(t); u^*(\cdot))}{\varepsilon} \geqslant 0, \tag{3.8}$$

where

$$u^\varepsilon(\cdot) = u(\cdot) I_{[t,t+\varepsilon)}(\cdot) + u^*(\cdot) I_{[t+\varepsilon,T]}(\cdot). \tag{3.9}$$

In this case, $X^*(\cdot)$ and $u^*(\cdot)$ are called an *open-loop equilibrium state process* and an *open-loop equilibrium control*, respectively.

We refer to (3.8) as a *local optimality condition* at $t \in [0,T)$. One sees that if $(X^*(\cdot), u^*(\cdot))$ is an open-loop equilibrium pair of Problem (LQ) for the initial state x, then along the open-loop equilibrium state $X^*(\cdot)$, the open-loop equilibrium control $u^*(\cdot)$ stays locally optimal. On the other hand, since $(X^*(\cdot), u^*(\cdot))$ is a fixed state-control pair of (1.1) on $[0,T]$, in which the conditional expectation terms are absent, the above defined open-loop equilibrium pair is time-consistent. Note that if we consider the general state equation (3.1) in which some conditional expectation terms appear, we do not know if one can directly define time-consistent state-control pairs. This is why for open-loop equilibrium solutions of Problem (LQ), we only consider (1.1). The following result, in some sense, is an extension of a relevant result found in [15].

Proposition 3.4. *Let* (LQ1)–(LQ2) *hold. Suppose* $(X^*(\cdot), u^*(\cdot)) \in \mathcal{X}[0,T] \times \mathcal{U}[0,T]$ *is a state-control pair starting from initial state* x. *For each* $t \in [0,T)$, *let* $(Y(\cdot\,,t), Z(\cdot\,,t))$ *be the adapted solution of the following BSDE:*

$$\begin{cases} dY(s,t) = -\Big\{A(s)^T Y(s,t) + C(s)^T Z(s,t) + Q(s,t)X^* \\ \qquad\qquad +\bar{Q}(s,t)\mathbb{E}_t[X^*(s)]\Big\}ds + Z(s,t)dW(s), \qquad s \in [t,T], \\ Y(T,t) = G(t)X^*(T) + \bar{G}(t)\mathbb{E}_t[X^*(T)]. \end{cases} \tag{3.10}$$

Suppose $(s,t)\mapsto (Y(s,t),Z(s,t))$ *is continuous on* $0\leqslant t\leqslant s\leqslant T$ *and suppose*

$$u^*(t)=-\widehat{R}(t,t)^{-1}\big\{B(t)^TY(t,t)+D(t)^TZ(t,t)\big\},\quad t\in[0,T]. \tag{3.11}$$

Then $(X^*(\cdot),u^*(\cdot))$ *is an open-loop equilibrium pair of Problem* (LQ) *for initial state* x.

The above leads to the following FBSDE family (parameterized by $t\in[0,T)$):

$$\begin{cases} dX^*(s)=\big\{A(s)X^*(s)+B(s)u^*(s)\big\}ds\\ \qquad\qquad +\big\{C(s)X^*(s)+D(s)u^*(s)\big\}dW(s),\quad 0\leqslant s\leqslant T,\\ dY(s,t)=-\big\{A(s)^TY(s,t)+C(s)^TZ(s,t)+Q(s,t)X^*(s)\\ \qquad\qquad +\bar{Q}(s,t)\mathbb{E}_t[X^*(s)]\big\}ds+Z(s,t)dW(s),\qquad 0\leqslant t\leqslant s\leqslant T,\\ X^*(0)=x,\qquad Y(T,t)=G(t)X^*(T)+\bar{G}(t)\mathbb{E}_t[X^*(T)],\\ \widehat{R}(t,t)u^*(t)+B(t)^TY(t,t)+D(t)^TZ(t,t)=0,\quad t\in[0,T]. \end{cases} \tag{3.12}$$

Inspired by [22], we suppose

$$Y(s,t)=P(s,t)X^*(s)+\bar{P}(s,t)\mathbb{E}_t[X^*(s)],\qquad s\in[t,T],$$

for some deterministic functions $P(\cdot\,,\cdot)$ and $\bar{P}(\cdot\,,\cdot)$. Then the above will be true if we let $P(s,t)$ and $\widehat{P}(s,t)\equiv P(s,t)+\bar{P}(s,t)$ be the solutions to the following coupled Riccati equations:

$$\begin{cases} P_s(s,t)+P(s,t)A(s)+A(s)^TP(s,t)+C(s)^TP(s,t)C(s)+Q(s,t)\\ \quad -[P(s,t)B(s)+C(s)^TP(s,t)D(s)][\widehat{R}(s,s)+D(s)^TP(s,s)D(s)]^{-1}\\ \quad\cdot[B(s)^TP(s,s)+D(s)^TP(s,s)C(s)]=0,\\ \widehat{P}_s(s,t)+\widehat{P}(s,t)A(s)+A(s)^T\widehat{P}(s,t)+C(s)^TP(s,t)C(s)+\widehat{Q}(s,t)\\ \quad -[\widehat{P}(s,t)B(s)+C(s)^TP(s,t)D(s)][\widehat{R}(s,s)+D(s)^TP(s,s)D(s)]^{-1}\\ \quad\cdot[B(s)^T\widehat{P}(s,s)+D(s)^TP(s,s)C(s)]=0,\qquad s\in[t,T],\\ P(T,t)=G(t),\qquad \widehat{P}(T,t)=\widehat{G}(t). \end{cases} \tag{3.13}$$

To summarize the above, we state the following result.

Theorem 3.5. *Let* (LQ1)–(LQ2) *hold. Suppose Riccati equation system* (3.13) *admits a unique solution* $(P(\cdot\,,\cdot),\widehat{P}(\cdot\,,\cdot))$ *which is continuous in both variables. Then an open-loop equilibrium control* $u^*(\cdot)\in\mathcal{U}[0,T]$ *exists and it admits the following closed-loop representation:*

$$u^*(t)=-\big[\widehat{R}(t,t)+D(t)^TP(t,t)D(t)\big]^{-1}\Big[B(t)^T\widehat{P}(t,t)+D(t)^TP(t,t)C(t)\Big]X^*(t),$$

where $X^*(\cdot)$ *is the state under control* $u^*(\cdot)$.

From the above, we see that the existence of an open-loop equilibrium control is guaranteed by the solvability of Riccati equation system (3.13).

We now make a couple of comments on this.

The advantages: The approach is direct and the derivation of equilibrium pair is not very complicated. Moreover, the open-loop equilibrium control $u^*(\cdot)$ admits a closed-loop representation.

The disadvantages: The Riccati equations in (3.13) do not have symmetry structure. Therefore the solutions $P(\cdot\,,\cdot)$ and $\widehat{P}(\cdot\,,\cdot)$ of the system are not necessarily symmetric. This leads to some difficulties in establish the well-posedness of the system.

3.2. Close-loop equilibrium strategy. In this subsection, we introduce closed-loop equilibrium strategies. We recall denote that any partition of $[0,T]$ is denoted by Δ:

$$\Delta = \big\{t_k \bigm| 0 \leqslant k \leqslant N\big\} \equiv \big\{0 = t_0 < t_1 < t_2 < \cdots < t_{N-1} < t_N = T\big\},$$

with N being some natural number, and define its *mesh size* by the following:

$$\|\Delta\| = \max_{0\leqslant k\leqslant N-1}(t_{k+1} - t_k).$$

For the above Δ, we define

$$\begin{aligned} J_k^\Delta(X(\cdot),u(\cdot)) = \mathbb{E}_{t_k}\Big\{ \int_{t_k}^T \big[\langle\, Q(s,t_k)X(s), X(s)\,\rangle + \langle\, \bar{Q}(s,t_k)\mathbb{E}_{t_k}[X(s)], \mathbb{E}_{t_k}[X(s)]\,\rangle \\ + \langle\, R(s,t_k)u(s), u(s)\,\rangle + \langle\, \bar{R}(s,t_k)\mathbb{E}_{t_k}[u(s)], \mathbb{E}_{t_k}[u(s)]\,\rangle \big] ds \\ + \langle\, G(t_k)X(T), X(T)\,\rangle + \langle\, \bar{G}(t_k)\mathbb{E}_{t_k}[X(T)], \mathbb{E}_{t_k}[X(T)]\,\rangle \Big\}, \end{aligned}$$

for any $(X(\cdot),u(\cdot)) \in \mathcal{X}[t_k,T]\times\mathcal{U}[t_k,T]$, $k = 0,1,2,\cdots,N-1$. In the above, $(X(\cdot),u(\cdot))$ does not have to be a state-control pair of the original control system.

Now, we introduce the following which is comparable with Definition 2,2.

Definition 3.6. Let $\Delta = \{0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T\}$ be a partition of $[0,T]$, and let $\Theta^\Delta, \widehat{\Theta}^\Delta : [0,T] \to \mathbb{R}^{m\times n}$ be two given maps depending on Δ.

(i) For any $x \in \mathbb{R}^n$ fixed, let $X^\Delta(\cdot) \equiv X^\Delta(\cdot\,;x)$ be the solution to the following linear MF-SDE:

$$\begin{cases} dX^\Delta(s) = \Big\{\big[A(s) - B(s)\Theta^\Delta(s)\big]X^\Delta(s) \\ \qquad\qquad + \Big[\bar{A}(s) + B(s)\big[\Theta^\Delta(s) - \widehat{\Theta}^\Delta(s)\big] - \bar{B}(s)\widehat{\Theta}^\Delta(s)\Big]\mathbb{E}_{\rho^\Delta(s)}[X^\Delta(s)]\Big\}ds \\ \qquad\qquad + \Big\{\big[C(s) - D(s)\Theta^\Delta(s)\big]X^\Delta(s) \\ \qquad\qquad + \Big[\bar{C}(s) + D(s)\big[\Theta^\Delta(s) - \widehat{\Theta}^\Delta(s)\big] - \bar{D}(s)\widehat{\Theta}^\Delta(s)\Big]\mathbb{E}_{\rho^\Delta(s)}[X^\Delta(s)]\Big\}dW(s), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s \in [0,T], \\ X^\Delta(0) = x, \end{cases}$$

where

$$\rho^\Delta(s) = \sum_{k=0}^{N-1} t_k I_{[t_k,t_{k+1})}(s), \qquad s \in [0,T],$$

and let $u^\Delta(\cdot) \equiv u^\Delta(\cdot\,;x)$ be defined by

$$u^\Delta(s) = -\Theta^\Delta(s)X^\Delta(s) + \big[\Theta^\Delta(s) - \widehat{\Theta}^\Delta(\cdot)\big]\mathbb{E}_{\rho^\Delta(s)}[X^\Delta(s)], \qquad s \in [0,T]. \tag{3.14}$$

The pair $(X^\Delta(\cdot),u^\Delta(\cdot))$ is called the *closed-loop pair* associated with Δ and $(\Theta^\Delta(\cdot),\widehat{\Theta}^\Delta(\cdot))$, starting from x.

(ii) For each $t_k \in \Delta$ and any $u_k(\cdot) \in \mathcal{U}[t_k,t_{k+1}]$, let $X_k(\cdot)$ be the solution to the following system:

$$\begin{cases} dX_k(s) = \Big\{A(s)X_k(s) + \bar{A}(s)\mathbb{E}_{t_k}[X_k(s)] + B(s)u_k(s) + \bar{B}(s)\mathbb{E}_{t_k}[u_k(s)]\Big\}ds \\ \qquad\qquad + \Big\{C(s)X_k(s) + \bar{C}(s)\mathbb{E}_{t_k}[X_k(s)] + D(s)u_k(s) + \bar{D}(s)\mathbb{E}_{t_k}[u_k(s)]\Big\}dW(s), \\ \qquad\qquad\qquad\qquad s \in [t_k,t_{k+1}], \\ X_k(t_k) = X^\Delta(t_k), \end{cases}$$

and $X^{\Delta}_{k+1}(\cdot)$ be the solution to the following:

$$\begin{cases} dX^{\Delta}_{k+1}(s)=\Big\{\big[A(s)-B(s)\Theta^{\Delta}(s)\big]X^{\Delta}_{k+1}(s) \\ \qquad\qquad +\Big[\bar{A}(s)+B(s)\big[\Theta^{\Delta}(s)-\widehat{\Theta}^{\Delta}(s)\big]-\bar{B}(s)\widehat{\Theta}^{\Delta}(s)\Big]\mathbb{E}_{\rho^{\Delta}(s)}[X^{\Delta}_{k+1}(s)]\Big\}ds \\ \qquad\qquad +\Big\{\big[C(s)-D(s)\Theta^{\Delta}(s)\big]X^{\Delta}_{k+1}(s) \\ \qquad\qquad +\Big[\bar{C}(s)+D(s)\big[\Theta^{\Delta}(s)-\widehat{\Theta}^{\Delta}(s)\big]-\bar{D}(s)\widehat{\Theta}^{\Delta}(s)\Big]\mathbb{E}_{\rho^{\Delta}(s)}[X^{\Delta}_{k+1}(s)]\Big\}dW(s), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \in[t_{k+1},T], \\ X^{\Delta}_{k+1}(t_{k+1})=X_k(t_{k+1}). \end{cases}$$

Denote

$$\begin{cases} X_k(\cdot)\oplus X^{\Delta}(\cdot)\equiv X_k(\cdot)I_{[t_k,t_{k+1})}(\cdot)+X^{\Delta}_{k+1}(\cdot)I_{[t_{k+1},T]}(\cdot), \\ u_k(\cdot)\oplus u^{\Delta}(\cdot)=u_k(\cdot)I_{[t_k,t_{k+1})}(\cdot) \\ \qquad -\big\{\Theta^{\Delta}(\cdot)X^{\Delta}_{k+1}(\cdot)+\big[\Theta^{\Delta}(\cdot)-\widehat{\Theta}^{\Delta}(\cdot)\big]\mathbb{E}_{\rho^{\Delta}(\cdot)}[X^{\Delta}_{k+1}(\cdot)]\big\}I_{[t_{k+1},T]}(\cdot). \end{cases}$$

We call $(X_k(\cdot)\oplus X^{\Delta}(\cdot),u_k(\cdot)\oplus u^{\Delta}(\cdot))$ a *local variation* of $(X^{\Delta}(\cdot),u^{\Delta}(\cdot))$ on $[t_k,t_{k+1}]$. Suppose the following *local optimality condition* holds:

$$J^{\Delta}_k\big(X^{\Delta}_k(\cdot),u^{\Delta}_k(\cdot)\big)\leqslant J^{\Delta}_k\big(X_k(\cdot)\oplus X^{\Delta}(\cdot),u_k(\cdot)\oplus u^{\Delta}(\cdot)\big),\quad \forall u_k(\cdot)\in\mathcal{U}[t_k,t_{k+1}].$$

Then we call $(\Theta^{\Delta}(\cdot),\widehat{\Theta}^{\Delta}(\cdot))$ a *closed-loop Δ-equilibrium strategy* of Problem (MF-LQ), and call $\big(X^{\Delta}(\cdot\,;x),u^{\Delta}(\cdot\,;x)\big)$ a *closed-loop Δ-equilibrium pair* of Problem (MF-LQ) for the initial state x.

(iii) If the following holds:

$$\lim_{\|\Delta\|\to 0}\Big[\|\Theta^{\Delta}(\cdot)-\Theta(\cdot)\|_{C([0,T];\mathbb{R}^{m\times n})}+\|\widehat{\Theta}^{\Delta}(\cdot)-\widehat{\Theta}(\cdot)\|_{C([0,T];\mathbb{R}^{m\times n})}\Big]=0,$$

for some $\Theta,\widehat{\Theta}\in C([0,T];\mathbb{R}^{m\times n})$, then $(\Theta(\cdot),\widehat{\Theta}(\cdot))$ is called a *closed-loop equilibrium strategy* of Problem (MF-LQ). For any $(t,x)\in\mathcal{D}$, let $\widehat{X}^*(\cdot)\equiv\widehat{X}^*(\cdot\,;t,x)$ be the solution to the following system:

$$\begin{cases} d\widehat{X}^*(s)=\big[A(s)-B(s)\widehat{\Theta}(s)\big]\widehat{X}^*(s)ds+\big[C(s)-D(s)\widehat{\Theta}(s)\big]\widehat{X}^*(s)dW(s), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s\in[t,T], \\ \widehat{X}^*(t)=x, \end{cases}$$

and define $\widehat{u}^*(\cdot)\equiv\widehat{u}^*(\cdot\,;t,x)$ as follows:

$$\widehat{u}^*(s)=-\widehat{\Theta}(s)\widehat{X}^*(s),\qquad s\in[t,T].$$

Then $(t,x)\mapsto(\widehat{X}^*(\cdot\,;t,x),\widehat{u}^*(\cdot\,;t,x))$ is called a *closed-loop equilibrium pair flow* of Problem (MF-LQ). Further,

$$\widehat{V}(t,x)=\widetilde{J}(t,x;\widehat{X}^*(\cdot\,;t,x),\widehat{u}^*(\cdot\,;t,x)),\qquad (t,x)\in\mathcal{D}$$

is called a *closed-loop equilibrium value function* of Problem (MF-LQ).

We point out that $(\Theta^\Delta(\cdot),\widehat{\Theta}^\Delta(\cdot))$ and $(\Theta(\cdot),\widehat{\Theta}(\cdot))$ are independent of the initial state $x\in\mathbb{R}^n$. Let us now state the main result of this section.

Theorem 3.7. *Let* (LQ1)–(LQ3) *hold. Then there exists a unique pair* $(\Gamma,\widehat{\Gamma})$ *of* $\mathbb{S}^n$*-valued functions solving the following system of equations:*

$$\begin{cases}\Gamma_s(s,t)+\Gamma(s,t)\big[A(s)-B(s)\widehat{\Theta}(s)\big]+\big[A(s)-B(s)\widehat{\Theta}(s)\big]^T\Gamma(s,t)+Q(s,t)\\ \quad+\big[C(s)-D(s)\widehat{\Theta}(s)\big]^T\Gamma(s,t)\big[C(s)-D(s)\widehat{\Theta}(s)\big]+\widehat{\Theta}(s)^TR(s,t)\widehat{\Theta}(s)=0,\\ \widehat{\Gamma}_s(s,t)+\widehat{\Gamma}(s,t)\big[A(s)-B(s)\widehat{\Theta}(s)\big]+\big[A(s)-B(s)\widehat{\Theta}(s)\big]^T\widehat{\Gamma}(s,t)+\widehat{Q}(s,t)\\ \quad+\big[C(s)-D(s)\widehat{\Theta}(s)\big]^T\Gamma(s,t)\big[C(s)-D(s)\widehat{\Theta}(s)\big]+\widehat{\Theta}(s)^T\widehat{R}(s,t)\widehat{\Theta}(s)=0,\\ \qquad\qquad 0\leqslant t\leqslant s\leqslant T,\\ \Gamma(T,t)=G(t),\qquad \widehat{\Gamma}(T,t)=\widehat{G}(t),\qquad 0\leqslant t\leqslant T,\end{cases}$$

where $\widehat{\Theta}(\cdot)$ *is given by the following:*

$$\widehat{\Theta}(s)=\big[\widehat{R}(s,s)+D(s)^T\Gamma(s,s)D(s)\big]^{-1}\big[B(s)^T\widehat{\Gamma}(s,s)+D(s)^T\Gamma(s,s)C(s)\big],\quad s\in[0,T].$$

The closed-loop equilibrium state process $X^*(\cdot)$ *is the solution to the following closed-loop system:*

$$\begin{cases}dX^*(s)=\big[A(s)-B(s)\widehat{\Theta}(s)\big]X^*(s)ds+\big[C(s)-D(s)\widehat{\Theta}(s)\big]X^*(s)dW(s),\\ \qquad\qquad s\in[0,T],\\ X^*(0)=x,\end{cases}$$

the closed-loop equilibrium control admits the following representation:

$$u^*(s)=-\widehat{\Theta}(s)X^*(s),\quad s\in[0,T],\tag{3.15}$$

and the closed-loop equilibrium value function is given by the following:

$$\widehat{V}(t,x)=\langle\,\widehat{\Gamma}(t,t)x,x\,\rangle,\qquad\forall(t,x)\in\mathcal{D}.\tag{3.16}$$

Note that in Theorem 3.6, the Riccati equations for $\Gamma(\cdot\,,\cdot)$ and $\widehat{\Gamma}(\cdot\,,\cdot)$ are different: $(Q(\cdot\,,\cdot),R(\cdot\,,\cdot),G(\cdot))$ appears in the former and $(\widehat{Q}(\cdot\,,\cdot),\widehat{R}(\cdot\,,\cdot),\widehat{G}(\cdot))$ appears in the later. Also, we see that the system is fully coupled.

Directly comparing the results of this subsection with those in the previous subsection, we see that the open-loop and closed-loop equilibrium solutions are different for Problem (LQ). The results coincide when the problem is reduced to classical LQ problems.

The proof is lengthy and technical. However, the main idea is similar to that in Section 2, based on a careful analysis of multi-person differential games. For details, see [37].

Acknowledgements. This work is supported in part by NSF Grant DMS-1007514.

References

[1] M. Allais, *Le compaortment del'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine*, Econometrica, **21** (1953), 503–546.

[2] D. Bernoulli, *Specimen theoriae novae de mensura sortis, Commentarii Academiae Scientiatatum Imperalas Petropolitanae*, **5** (1738), 175–192: Translated as Exposition of a New Theory on the Measurement of Risk, Econometrica **22** (1954), 23 – 26.

[3] T. Björk and A. Murgoci, *A general theory of Markovian time inconsistent stochasitc control problem*, working paper.

[4] T. Björk, A. Murgoci, and X. Y. Zhou, *Mean varaiance portfolio optimization with state dependent risk aversion*, Math. Finance, in press.

[5] G. Choquet, *Theory of capacities*, Ann. Inst. Fourier, **5**, 131–296.

[6] B. de Finetti, *La prévision: Ses lois logiques, ses sources subjectives*, Annals de l'Institute Henri Poincare, **7** (1937), 1–68.

[7] D. Denneberg, Non-Additive Measure and Integral, Kluwer Academic Publishers, Dordrecht, 1994.

[8] D. Duffie and L. G. Epstein, *Stochastic differential utility*, Econometrica, **60** (1992), 353–394.

[9] I. Ekeland and A. Lazrak, *The golden rule when preferences are time inconsistent*, Math. Finan. Econ., **4** (2010), 29–55.

[10] I. Ekeland and T. A. Pirvu, *Investment and consumption without commitment*, Math. Finan. Econ., **2** (2008), 57–86.

[11] I. Ekeland, O. Mbodji, and T. A. Pirvu, *Time-Consistent Portfolio Management*, SIAM J. Financial Math., **3** (2012), 1–32.

[12] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd Edition, Springer-Verlag, New York, 2006.

[13] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, **7** (1997), 1–71.

[14] D. Ellsberg, *Risk ambiguity and the Savage axioms*, Quartely J. Economics, **75** (1961), 643–649.

[15] Y. Hu, H. Jin, and X. Y. Zhou, *Time-inconsistent stochastic linear-quadratic control*, arXiv:1111.0818v1, 2011.

[16] D. Hume, *A Treatise of Human Nature*, First Edition, 1739; Reprint, Oxford Univ. Press, New York, 1978.

[17] H. Jin and X. Y. Zhou, *Behavioral portfolio selection in continuous time*, Math Finance, **18** (2008), 385–426.

[18] N. V. Krylov, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, Reidel, Dordrecht, 1987.

[19] D. Laibson, *Golden eggs and hyperbolic discounting*, Quarterly J. Econ., **112** (1997), 443–477.

[20] D. Kahneman and A. Tversky, *Prospect theory: An analysis of decision under risk*, Econometrica, **47** (1979), 263–291.

[21] J. Ma, P. Protter, and J. Yong, *Solving forward-backward stochastic differential equations explicitly — a four-step scheme*, Probab. Theory & Related Fields, **98** (1994), 339–359.

[22] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math., Vol. 1702, Springer-Verlag, 1999.

[23] J. Marin-Solano, and J. Navas, *Consumption and portfolio rules for time-inconsistnet investors*, European J. Operational Research, **201** (2010), 860–872.

[24] J. Marin-Solano and E. V. Shevkoplyas, *Non-constant discounting and differential games with random time horizon*, Automatica, **47** (2011), 2626–2638.

[25] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton Univ. Press, 1944.

[26] I. Palacios-Huerta, *Time-inconsistent preferences in Adam Smith and Davis Hume*, History of Political Economy, **35** (2003), 241–268.

[27] S. Peng, *A generalized dynamic programming principle and Hmilton-Jacobi-Bellman equation*, Stoch. Stoch. Rep., **38** (1992), 119–134.

[28] ———, *Backward stochastic differential equations and stochastic optimizations*, Topics in Stochastic Analysis, J. Yan, S. Peng, S. Fang, and L. Wu, eds., Science Press, Beijing, 1997 (in Chinese).

[29] R. A. Pollak, *Consistent planning, Review of Economic Studies*, **35** (1968), 185–199.

[30] R. H. Strotz, *Myopia and inconsistency in dynamic utility maximization*, Review of Econ. Studies, **23** (1955), 165–180.

[31] F. P. Ramsey, *Truth and probability*, In The Foundations of Mathematics and Other Logical Essays. R. B. Braithwaite and F. Plumpton (Eds.) London: K. Paul, Trench, Truber and Co.

[32] L. J. Savage, *The Foundations of Statistics*, Wiley, New York, 1954.

[33] S. S. Wang, V. R. Young, and H. H. Panjer, *Axiomatic characterization of insurance prices*, Insurance: Math & Econ., **21** (1997), 173–183.

[34] J. Yong, *A deterministic linear quadratic time-inconsistent optimal control problem*, Math. Control & Related Fields, **1** (2011), 83–118.

[35] ———, *Deterministic time-inconsistent optimal control problems — An essentially cooperative approach*, Acta Appl. Math. Sinica, **28** (2012), 1–20.

[36] ———, *Time-inconsistent optimal control problems and the equilibrium HJB equation*, Math. Control & Related Fields, **2** (2012), 271–329.

[37] ———, *A linear-quadratic optimal control problem for mean-field stochastic differential equations*, SIAM J. Control Optim., **51** (2013), 2809–2838.

[38] _____, *Linear-quadratic optimal control problems for mean-field stochastic differential equations — time-consistent solutions*, arXiv:1304.3964v1.

[39] J. Yong, and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.

Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
E-mail: jiongmin.yong@ucf.edu

17. Mathematics in Science and Technology

Mathematical models and numerical methods for Bose-Einstein condensation

Weizhu Bao

Abstract. The achievement of Bose-Einstein condensation (BEC) in ultracold vapors of alkali atoms has given enormous impulse to the theoretical and experimental study of dilute atomic gases in condensed quantum states inside magnetic traps and optical lattices. This article offers a short survey on mathematical models and theories as well as numerical methods for BEC based on the mean field theory. We start with the Gross-Pitaevskii equation (GPE) in three dimensions (3D) for modeling one-component BEC of the weakly interacting bosons, scale it to obtain a three-parameter model and show how to reduce it to two dimensions (2D) and one dimension (1D) GPEs in certain limiting regimes. Mathematical theories and numerical methods for ground states and dynamics of BEC are provided. Extensions to GPE with an angular momentum rotation term for a rotating BEC, to GPE with long-range anisotropic dipole-dipole interaction for a dipolar BEC and to coupled GPEs for spin-orbit coupled BECs are discussed. Finally, some conclusions are drawn and future research perspectives are discussed.

Mathematics Subject Classification (2010). Primary 35Q55; Secondary 70F10.

Keywords. Bose-Einstein condensation, Gross-Pitaevskii equation, nonlinear Schrödinger equation, ground state, dynamics, numerical methods.

1. Introduction

The achievement of Bose-Einstein condensation (BEC) of dilute gases in 1995 [3, 28, 39] marked the beginning of a new era in atomic, molecular and optical (AMO) physics and quantum optics. In fact, the phenomenon known as BEC was predicted by Einstein in 1924 [40, 41] based on the ideas of Bose [27] concerning photons: In a system of bosons obeying Bose statistics under the assumption that it is in equilibrium at temperature T and chemical potential μ, Einstein [40, 41] derived the so-called Bose-Einstein distribution (or Bose-Einstein statistics), in the grand canonical ensemble, for the mean occupation of the jth energy state as

$$n_j = \frac{1}{e^{(\varepsilon_j-\mu)/k_BT}-1} := f(\varepsilon_j), \qquad j = 0, 1, \dots , \tag{1.1}$$

where $\varepsilon_j > \mu$ is the energy of the jth state, n_j is the number of particles in state j, k_B is the Boltzmann constant. The mean total number of particles is given as $N(T, \mu) = \sum_{j=0}^{\infty} f(\varepsilon_j)$,

Proceedings of the International Congress of Mathematicians, Seoul, 2014

and the mean total energy is given as $E(T,\mu) = \sum_{j=0}^{\infty} \varepsilon_j f(\varepsilon_j)$. From the above distribution, Einstein [40, 41] predicted that there should be a critical temperature T_c below which a finite fraction of all the particles "condense" into the same one-particle state.

Einstein's original prediction was for a noninteracting gas and did not receive much attention in a long time. After the observation of superfluidity in liquid ^{4}He below the λ temperature (2.17K) in 1938, London [61] suggested that despite the strong interatomic interactions BEC was indeed occurring in this system and was responsible for the superfluid properties. This suggestion has stood the test of time and is the basis for our modern understanding of the properties of the superfluid phase. By combining laser cooling and evaporative cooling, in 1995 BEC was realized in a system that is about as different as possible from ^{4}He, namely, dilute atomic alkali gases trapped by magnetic fields and over the last two decades these systems have been the subject of an explosion of research, both experimentally and theoretically. Perhaps the single aspect of BEC systems that makes them most fascinating is best illustrated by the cover of *Science* magazine of December 22, 1995, in which the Bose condensate is declared "**molecule of the year**" and pictured as a platoon of soldiers marching in lock-step: every atom in the condensate must behave in exactly the same way, and this has the consequence, *inter alia*, that effects which are so small as to be essentially invisible at the level of single atom may be spectacularly amplified. Most BEC experiments reach quantum degeneracy between 50 nK and 2 μK, at densities between 10^{11} and 10^{15} cm^{-3}. The largest condensates are of 100 million atoms for sodium, and a billion for hydrogen; the smallest are just a few hundred atoms. Depending on the magnetic trap, the shape of the condensate is either approximately round, with a diameter of 10–15 μm, or cigar-shaped with about 15 μm in diameter and 300 μm in length. The full cooling cycle that produce a condensate may take from a few seconds to as long as several minutes [37, 52]. For better understanding of the long history towards the BEC and its physical study, we refer to the Nobel lectures [37, 52] and several review papers in physics [38, 56, 65, 67].

The experimental advances in BEC [3, 28, 39] have spurred great excitement in the AMO community and condense matter community as well as computational and applied mathematics community. Since 1995, numerous efforts have been devoted to the studies of ultracold atomic gases and various kinds of condensates of dilute gases have been produced for both bosonic particles and fermionic particles [38, 43, 56]. In this rapidly growing research area, mathematical models and analysis as well as numerical simulation have been playing an important role in understanding the theoretical part of BEC and predicting and guiding the experiments. The goal of this paper is to offer a short survey on mathematical models and theories as well as numerical methods for BEC based on the Gross-Pitaevskii equation (GPE) [7, 46, 65–67]. The paper is organized as follows. In section 2, we present the GPE for BEC based on the mean field approximation. Ground states and their computations are discussed in section 3, and dynamics and its computation are presented in section 4. Extensions to rotating BEC, dipolar BEC and spin-orbit-coupled BEC are presented in section 5. Finally, some conclusions and perspectives are drawn in section 6.

2. The Gross-Pitaevskii equation

In this section, we will present the GPE for modeling BEC based on the mean field approximation [7, 46, 65–67], its nondimensionalization and dimension reduction to lower

dimensions.

2.1. Mean field approximation. For a BEC of ultracold dilute gas with N identical bosons confined in an external trap, only binary interaction is important, then the many-body Hamiltonian for it can be written as [56, 58]

$$H_N = \sum_{j=1}^{N} \left(-\frac{\hbar^2}{2m}\Delta_j + V(\mathbf{x}_j) \right) + \sum_{1\le j<k\le N} V_{\rm int}(\mathbf{x}_j - \mathbf{x}_k), \tag{2.1}$$

where $\mathbf{x}_j \in \mathbb{R}^3$ denotes the position of the jth particle for $j = 1, \ldots, N$, m is the mass of a boson, $\hbar$ is the Planck constant, $\Delta_j = \nabla_j^2$ is the Laplace operator with respect to $\mathbf{x}_j$, $V(\mathbf{x}_j)$ is the external trapping potential, and $V_{\rm int}(\mathbf{x}_j - \mathbf{x}_k)$ denotes the inter-atomic two body interaction. Denote the complex-valued wave function $\Psi_N := \Psi_N(\mathbf{x}_1, \ldots, \mathbf{x}_N, t) \in L^2(\mathbb{R}^{3N} \times \mathbb{R})$ for the N particles in the BEC, which is symmetric with respect to any permutation of the positions $\mathbf{x}_j$ $(1 \le j \le N)$, then the total energy is given as

$$E_{\rm total}(\Psi_N) = (\Psi_N, H_N\Psi_N) := \int_{\mathbb{R}^{3N}} \overline{\Psi}_N H_N \Psi_N \, d\mathbf{x}_1 \ldots d\mathbf{x}_N, \tag{2.2}$$

where $\overline{f}$, $\mathrm{Re}(f)$ and $\mathrm{Im}(f)$ denote the complex conjugate, real part and imaginary part of f, respectively, and the evolution of the system is described by the time-dependent linear Schrödinger equation

$$i\hbar\partial_t \Psi_N(\mathbf{x}_1, \ldots, \mathbf{x}_N, t) = \frac{\delta E_{\rm total}(\Psi_N)}{\delta \overline{\Psi}_N} = H_N \Psi_N(\mathbf{x}_1, \ldots, \mathbf{x}_N, t), \tag{2.3}$$

where $i = \sqrt{-1}$ denotes the imaginary unit and t is time.

For a BEC, all particles are in the same quantum state and we can formally take the Hartree ansatz [7, 42, 46, 58, 59, 65–67]

$$\Psi_N(\mathbf{x}_1, \ldots, \mathbf{x}_N, t) \approx \prod_{j=1}^{N} \psi(\mathbf{x}_j, t), \tag{2.4}$$

with the normalization for the single-particle wave function $\psi := \psi(\mathbf{x}, t)$ as

$$\|\psi(\cdot, t)\|^2 := \int_{\mathbb{R}^3} |\psi(\mathbf{x}, t)|^2 \, d\mathbf{x} = 1, \tag{2.5}$$

where $\mathbf{x} = (x, y, z)^T \in \mathbb{R}^3$ is the Cartesian coordinate in three dimensions (3D). Due to that the BEC gas is dilute and the temperature is below the critical temperature T_c, i.e. a weakly interacting gas, the binary interaction $V_{\rm int}$ is well approximated by the effective contact interacting potential [65–67]:

$$V_{\rm int}(\mathbf{x}_j - \mathbf{x}_k) = g\,\delta(\mathbf{x}_j - \mathbf{x}_k), \tag{2.6}$$

where $\delta(\cdot)$ is the Dirac distribution and the constant $g = \frac{4\pi\hbar^2 a_s}{m}$ with a_s the s-wave scattering length of the bosons (positive for repulsive interaction and negative for attractive interaction, which is much smaller than the average distance between the particles). Plugging

(2.4) into (2.2), noticing (2.1) and (2.6), and keeping only the two-body interaction, we obtain $E_{\rm total}(\Psi_N) \approx N\,E(\psi)$ with the Gross-Pitaevskii (GP) energy (or energy per particle) defined as [46, 58, 59, 65–67]

$$E(\psi) = \int_{\mathbb{R}^3} \left[\frac{\hbar^2}{2m} |\nabla \psi(\mathbf{x},t)|^2 + V(\mathbf{x}) |\psi(\mathbf{x},t)|^2 + \frac{Ng}{2} |\psi(\mathbf{x},t)|^4 \right] d\mathbf{x}. \tag{2.7}$$

The dynamics of the BEC will be governed by the following nonlinear Schrödinger equation (NLSE) with cubic nonlinearity, known as the Gross-Pitaevskii equation (GPE) [7, 42, 46, 58, 59, 65–67]:

$$i\hbar \partial_t \psi = \frac{\delta E(\psi)}{\delta \overline{\psi}} = \left[-\frac{\hbar^2}{2m} \nabla^2 + V(\mathbf{x}) + Ng|\psi|^2 \right] \psi, \quad \mathbf{x} \in \mathbb{R}^3, \ \ t > 0. \tag{2.8}$$

In most BEC experiments, the trapping potential has been taken as the harmonic oscillator potential [3, 7, 28, 39, 67]

$$V(\mathbf{x}) = \frac{m}{2} \left(\omega_x^2 x^2 + \omega_y^2 y^2 + \omega_z^2 z^2 \right), \qquad \mathbf{x} = (x,y,z)^T \in \mathbb{R}^3, \tag{2.9}$$

where ω_x, ω_y and ω_z are the trap frequencies in x-, y- and z-direction, respectively. Without loss of generality, we assume that $\omega_x \le \omega_y \le \omega_z$ throughout the paper. For other trapping potentials used in BEC experiments, such as box potential, double-well potential and optical lattice potential, we refer to [7, 26, 65–68] and references therein.

The derivation of the GPE (2.8) from the linear Schrödinger equation (2.3) for a BEC (or a system of N identical particles) based on mean field approximation – dimension reduction – was formally obtained by Pitaevskii [66] and Gross [46] independently in 1960s. Since the first experimental observation of BEC in 1995, much attention has been paid to provide mathematical justification for the derivation when N is large enough: For ground states, Lieb et al. [58, 59] proved rigorously that the GP energy (2.7) approximates the energy of the many-body system correctly in the mean field regime; and for dynamics, Yau et al. [42] established the validity of the GPE (2.8) as an approximation for (2.3), which inspired great interests in the study on dynamics for such many body system recently [35, 36, 54]. The above GPE (2.8) is a very simple equation, which is very convenient for mathematical analysis and numerical calculations, and in the case of the BEC alkali gases, appears to give a rather good quantitative description of the behavior in a large variety of experiments [7, 65–67]. It has become the fundamental mathematical model for studying theoretically the ground states and dynamics of BECs [7, 65–67].

2.2. Nondimensionalization. In order to study theoretically BECs, we nondimensionalize the GPE (2.8) with the harmonic trapping potential (2.9) under the normalization (2.5) and introduce [7, 65–67]

$$\tilde{t} = \frac{t}{t_s}, \quad \tilde{\mathbf{x}} = \frac{\mathbf{x}}{x_s}, \quad \tilde{\psi}\left(\tilde{\mathbf{x}}, \tilde{t}\right) = x_s^{3/2} \psi\left(\mathbf{x}, t\right), \quad \tilde{E}(\tilde{\psi}) = \frac{E(\psi)}{E_s}, \tag{2.10}$$

where $t_s = \frac{1}{\omega_x}$, $x_s = \sqrt{\frac{\hbar}{m\omega_x}}$ and $E_s = \hbar\omega_x$ are the scaling parameters of dimensionless time, length and energy units, respectively. Plugging (2.10) into (2.8), multiplying by $t_s^2/m x_s^{1/2}$, and then removing all $\tilde{}$, we obtain the following dimensionless GPE under the

normalization (2.5) in 3D [7, 65–67]:

$$i\partial_t\psi(\mathbf{x},t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + \kappa|\psi(\mathbf{x},t)|^2\right]\psi(\mathbf{x},t), \quad \mathbf{x}\in\mathbb{R}^3, \ \ t>0, \tag{2.11}$$

where $\kappa = \frac{4\pi N a_s}{x_s}$ is the dimensionless interaction constant, the dimensionless trapping potential is given as [7, 65–67]

$$V(\mathbf{x}) = \frac{1}{2}\left(x^2+\gamma_y^2y^2+\gamma_z^2z^2\right), \ \mathbf{x}\in\mathbb{R}^3, \ \text{ with } \gamma_y = \frac{\omega_y}{\omega_x}\geq 1, \ \gamma_z = \frac{\omega_z}{\omega_x}\geq 1, \tag{2.12}$$

and dimensionless energy functional $E(\psi)$ is defined as [7, 65–67]

$$E(\psi) = \int_{\mathbb{R}^3}\left[\frac{1}{2}|\nabla\psi(\mathbf{x},t)|^2 + V(\mathbf{x})|\psi(\mathbf{x},t)|^2 + \frac{\kappa}{2}|\psi(\mathbf{x},t)|^4\right] d\mathbf{x}. \tag{2.13}$$

2.3. Dimension reduction. In many BEC experiments [3, 28, 39, 65–67], the trapping potential (2.12) is anisotropic, i.e. $\gamma_z \gg 1$ and/or $\gamma_y \gg 1$, and then the GPE in 3D can be further reduced to a GPE in two dimensions (2D) or one dimension (1D). Assume the initial data for the 3D GPE (2.11) is given as

$$\psi(\mathbf{x},0) = \psi_0(\mathbf{x}), \qquad \mathbf{x}\in\mathbb{R}^3, \tag{2.14}$$

and define the linear operator H as

$$H = -\frac{1}{2}\Delta + V(\mathbf{x}) = -\frac{1}{2}\nabla^2 + V(\mathbf{x}), \qquad \mathbf{x}\in\mathbb{R}^3. \tag{2.15}$$

When $\gamma_z \gg 1$ and $\gamma_y = O(1)$ ($\Leftrightarrow \omega_z \gg \omega_x$ and $\omega_y = O(\omega_x)$), i.e. disk-shaped condensate with strong confinement in the z-direction [3, 28, 39, 65, 67], then the linear operator H can be split as

$$H = -\frac{1}{2}\Delta_\perp + V_2(\mathbf{x}_\perp) - \frac{1}{2}\partial_{zz} + \frac{z^2}{2\varepsilon^4} := H_\perp + H_z^\varepsilon = H_\perp + \frac{1}{\varepsilon^2}H_{\tilde{z}}, \quad \mathbf{x}\in\mathbb{R}^3, \tag{2.16}$$

where $\mathbf{x}_\perp = (x,y)^T\in\mathbb{R}^2$, $\Delta_\perp = \partial_{xx}+\partial_{yy}$, $V_2(\mathbf{x}_\perp) = \frac{1}{2}(x^2+\gamma_y^2y^2)$, $H_\perp := -\frac{1}{2}\Delta_\perp + V_2(\mathbf{x}_\perp)$, $\varepsilon = 1/\sqrt{\gamma_z}$, $z = \varepsilon\tilde{z}$ and

$$H_z^\varepsilon := -\frac{1}{2}\partial_{zz} + \frac{z^2}{2\varepsilon^4} = \frac{1}{\varepsilon^2}\left[-\frac{1}{2}\partial_{\tilde{z}\tilde{z}} + \frac{\tilde{z}^2}{2}\right] := \frac{1}{\varepsilon^2}H_{\tilde{z}}, \qquad z,\tilde{z}\in\mathbb{R}. \tag{2.17}$$

For $H_{\tilde{z}}$ in (2.17), we know that the following linear eigenvalue problem

$$H_{\tilde{z}}\,\chi(\tilde{z}) = \left[-\frac{1}{2}\partial_{\tilde{z}\tilde{z}} + \frac{\tilde{z}^2}{2}\right]\chi(\tilde{z}) = \mu\,\chi(\tilde{z}), \qquad \tilde{z}\in\mathbb{R}, \tag{2.18}$$

with $\|\chi\|^2 := \int_{\mathbb{R}}|\chi(\tilde{z})|^2\,d\tilde{z} = 1$ admits distinct orthonormal eigenfunctions $\chi_k(\tilde{z})$ with corresponding eigenvalues μ_k for $k = 0,1,\ldots$ In fact, they form an orthonormal basis of $L^2(\mathbb{R})$ and can be chosen as [7, 14, 25, 65–67]

$$\mu_k = \frac{k+1}{2}, \quad \chi_k(\tilde{z}) = \frac{1}{\pi^{1/4}\sqrt{2^k\,k!}}e^{-\tilde{z}^2/2}\,H_k(\tilde{z}), \ \ \tilde{z}\in\mathbb{R}, \quad k = 0,1,2,\ldots, \tag{2.19}$$

with $H_k(\tilde{z})$ the standard Hermite polynomial of degree k. Thus $(\chi_k^\varepsilon(z), \mu_k^\varepsilon)$ for $k \geq 0$ are orthonormal eigenpairs to the operator H_z^ε with

$$\mu_k^\varepsilon = \frac{\mu_k}{\varepsilon^2} = \frac{k+1}{2\varepsilon^2}, \qquad \chi_k^\varepsilon(z) = \frac{1}{\sqrt{\varepsilon}}\chi_k(\tilde{z}) = \frac{1}{\sqrt{\varepsilon}}\chi_k\left(\frac{z}{\varepsilon}\right), \quad z \in \mathbb{R}. \tag{2.20}$$

For simplicity of notation, here we only consider "pure state" case in the strong confinement direction, especially the "ground state" case [7, 14, 25, 65–67]. Assuming that the initial data ψ_0 in (2.14) satisfies

$$\psi_0(\mathbf{x}) \approx \psi_2(\mathbf{x}_\perp)\chi_0^\varepsilon(z), \qquad \mathbf{x} \in \mathbb{R}^3, \qquad 0 < \varepsilon \ll 1, \tag{2.21}$$

noting the scale separation in (2.16), when $\varepsilon \to 0^+$, the solution ψ to the 3D GPE (2.11) can be well approximated as [7, 14, 25, 65–67]

$$\psi(\mathbf{x}, t) \approx \psi_2(\mathbf{x}_\perp, t)\ \chi_0^\varepsilon(z)\ e^{-i\,\mu_0^\varepsilon\, t}, \qquad \mathbf{x} \in \mathbb{R}^3, \quad t \geq 0. \tag{2.22}$$

Plugging (2.22) into (2.11) and then multiplying by $\chi_0^\varepsilon(z)\, e^{i\,\mu_0^\varepsilon\, t}$, integrating for z over $\mathbb{R}$, we obtain formally the GPE in 2D with $\psi_2 := \psi_2(\mathbf{x}_\perp, t)$ as [7, 14, 25, 65–67]

$$i\partial_t\psi_2 = \left[-\frac{1}{2}\Delta_\perp + V_2(\mathbf{x}_\perp) + \kappa\sqrt{\frac{\gamma_z}{2\pi}}\,|\psi_2|^2\right]\psi_2, \quad \mathbf{x}_\perp \in \mathbb{R}^2,\ \ t > 0. \tag{2.23}$$

The above dimension reduction from 3D to 2D is mathematically and rigorously justified in the very weak interaction regime [6, 25], i.e. $\kappa = O(\varepsilon) = O(1/\sqrt{\gamma_z})$ as $\varepsilon \to 0^+$. However, for the strong interaction regime, i.e. $\kappa = O(1)$ and $\varepsilon \to 0^+$, it is very challenging. The key difficulty is due to that the energy associated to the 2D GPE (2.23) is unbounded in this regime. Recently, by using a proper re-scaling, the dimension reduction is justified in this regime too [16].

Similarly, when $\gamma_z \gg 1$ and $\gamma_y \gg 1$ ($\Leftrightarrow \omega_z \gg \omega_x$ and $\omega_y \gg \omega_x$), i.e. cigar-shaped condensate with strong confinement in the (y, z)-plane [3, 28, 39, 65, 67], the 3D GPE (2.11) can be reduced to the following GPE in 1D as [7, 14, 65–67]

$$i\partial_t\psi_1(x,t) = \left[-\frac{1}{2}\partial_{xx} + \frac{x^2}{2} + \kappa\frac{\sqrt{\gamma_y\gamma_z}}{2\pi}\,|\psi_1(x,t)|^2\right]\psi_1(x,t), \ \ x \in \mathbb{R},\ t > 0. \tag{2.24}$$

Then the 3D GPE (2.11), 2D GPE (2.23) and 1D GPE (2.24) can be written in a unified way [7, 14, 65–67]

$$i\partial_t\psi(\mathbf{x},t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + \beta\,|\psi(\mathbf{x},t)|^2\right]\psi(\mathbf{x},t), \qquad \mathbf{x} \in \mathbb{R}^d, \quad t > 0, \tag{2.25}$$

where $\beta = \kappa$, $\kappa\sqrt{\gamma_z/2\pi}$ and $\kappa\sqrt{\gamma_y\gamma_z}/2\pi$ when $d = 3$, 2 and 1, respectively, and

$$V(\mathbf{x}) = \frac{1}{2}\begin{cases} x^2, & d = 1, \\ x^2 + \gamma_y^2 y^2, & d = 2, \\ x^2 + \gamma_y^2 y^2 + \gamma_z^2 z^2, & d = 3, \end{cases} \qquad \mathbf{x} \in \mathbb{R}^d. \tag{2.26}$$

This GPE conserves the normalization (or mass)

$$N(\psi(\cdot,t)) = \int_{\mathbb{R}^d} |\psi(\mathbf{x},t)|^2\, d\mathbf{x} \equiv \int_{\mathbb{R}^d} |\psi(\mathbf{x},0)|^2\, d\mathbf{x} = 1, \qquad t \geq 0, \tag{2.27}$$

and the energy per particle

$$E(\psi(\cdot,t)) = \int_{\mathbb{R}^d} \left[\frac{1}{2}|\nabla\psi|^2 + V(\mathbf{x})|\psi|^2 + \frac{\beta}{2}|\psi|^4\right] d\mathbf{x} \equiv E(\psi(\cdot,0)), \quad t \geq 0. \tag{2.28}$$

In fact, the energy functional $E(\psi)$ can be split into three parts as $E(\psi) = E_{\rm kin}(\psi) + E_{\rm pot}(\psi) + E_{\rm int}(\psi)$ with the kinetic energy $E_{\rm kin}(\psi)$, potential energy $E_{\rm pot}(\psi)$ and interaction energy $E_{\rm int}(\psi)$ defined as

$$E_{\rm kin}(\psi) = \int_{\mathbb{R}^d} \frac{1}{2}|\nabla\psi|^2 d\mathbf{x}, \quad E_{\rm int}(\psi) = \int_{\mathbb{R}^d} \frac{\beta}{2}|\psi|^4 d\mathbf{x}, \quad E_{\rm pot}(\psi) = \int_{\mathbb{R}^d} V(\mathbf{x})|\psi|^2 d\mathbf{x}.$$

3. Ground states

To find the stationary state of the GPE (2.25) for a BEC, we write [7, 12, 65–67]

$$\psi(\mathbf{x},t) = \phi(\mathbf{x})\, e^{-i\mu t}, \qquad \mathbf{x} \in \mathbb{R}^d, \quad t \geq 0, \tag{3.1}$$

where μ is the chemical potential of the condensate and $\phi(\mathbf{x})$ is a function independent of time. Substituting (3.1) into (2.25) gives the following for (μ, ϕ):

$$\mu\, \phi(\mathbf{x}) = -\frac{1}{2}\nabla^2\phi(\mathbf{x}) + V(\mathbf{x})\phi(\mathbf{x}) + \beta|\phi(\mathbf{x})|^2\phi(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^d, \tag{3.2}$$

under the normalization condition

$$\|\phi\|^2 := \int_{\mathbb{R}^d} |\phi(\mathbf{x})|^2 d\mathbf{x} = 1. \tag{3.3}$$

This is a nonlinear eigenvalue problem with a constraint and any eigenvalue μ can be computed from its corresponding eigenfunction $\phi(\mathbf{x})$ by [7, 12, 65–67]

$$\mu = \mu(\phi) = E(\phi) + \int_{\mathbb{R}^d} \frac{\beta}{2}|\phi(\mathbf{x})|^4 d\mathbf{x} = E(\phi) + E_{\rm int}(\phi). \tag{3.4}$$

The ground state of a BEC is usually defined as the minimizer of the following nonconvex (or constrained) minimization problem [7, 12]: Find $\phi_g \in S$ such that

$$E_g := E(\phi_g) = \min_{\phi\in S} E(\phi), \qquad \text{with } \mu_g := \mu(\phi_g) = E(\phi_g) + E_{\rm int}(\phi_g), \tag{3.5}$$

where $S = \{\phi \mid \|\phi\| = 1,\ E(\phi) < \infty\}$ is the unit sphere and μ_g is the corresponding chemical potential. It is easy to show that the ground state ϕ_g is an eigenfunction of the nonlinear eigenvalue problem (3.2) under the constraint (3.3), which is the Euler-Lagrangian equation of constrained minimization problem (3.5). Any eigenfunction of (3.2) whose energy is larger than that of the ground state is usually called excited states in the physics literatures.

3.1. Existence and uniqueness. Denote the best Sobolev constant C_b in 2D as

$$C_b := \inf_{0\neq f\in H^1(\mathbb{R}^2)} \frac{\|\nabla f\|^2_{L^2(\mathbb{R}^2)}\, \|f\|^2_{L^2(\mathbb{R}^2)}}{\|f\|^4_{L^4(\mathbb{R}^2)}}. \tag{3.6}$$

The best constant C_b can be attained at some H^1 function [7] and it is crucial in considering the existence of ground states in 2D. For existence and uniqueness of the ground state to (3.5), we have the following results.

Theorem 3.1 (Existence and uniqueness [7, 59]). *Suppose $V(\mathbf{x}) \geq 0$ ($\mathbf{x} \in \mathbb{R}^d$) in the energy functional (2.28) satisfies the confining condition $\lim_{|\mathbf{x}|\to\infty} V(\mathbf{x}) = \infty$, then there exists a ground state $\phi_g \in S$ for (3.5) if one of the following holds: (i) $d = 3$, $\beta \geq 0$; (ii) $d = 2$, $\beta > -C_b$; (iii) $d = 1$, for all $\beta \in \mathbb{R}$. Moreover, the ground state can be chosen as nonnegative $|\phi_g|$, and $\phi_g(\mathbf{x}) = e^{i\theta_0}|\phi_g(\mathbf{x})|$ for some constant $\theta_0 \in \mathbb{R}$. For $\beta \geq 0$, the nonnegative ground state ϕ_g is unique. If the potential $V(\mathbf{x}) \in L^2_{\rm loc}$, the nonnegative ground state is strictly positive. In contrast, there exists no ground state if one of the following holds: (i') $d = 3$, $\beta < 0$; (ii') $d = 2$, $\beta \leq -C_b$.*

For the ground state $\phi_g \in S$ of (3.5) with the harmonic potential (2.26), we have the following properties.

Theorem 3.2 (Virial identity [7, 67]). *The ground state $\phi_g \in S$ of (3.5) satisfies the following virial identity*

$$2E_{\rm kin}(\phi_g) - 2\, E_{\rm pot}(\phi_g) + d\, E_{\rm int}(\phi_g) = 0. \tag{3.7}$$

Theorem 3.3 (Symmetry [7, 59]). *Suppose $\gamma_y = \gamma_z = 1$ in (2.26), i.e. the harmonic trapping potential $V(\mathbf{x})$ is radially/spherically symmetric in 2D/3D and monotone increasing, then the positive ground state $\phi_g \in S$ of (3.5) must be radially/spherically symmetric in 2D/3D and monotonically decreasing, i.e. $\phi_g(\mathbf{x}) = \phi_g(r)$ with $r = |\mathbf{x}|$ for $\mathbf{x} \in \mathbb{R}^d$.*

Theorem 3.4 (Decay at far-field [7]). *When $\beta \geq 0$, for any $\nu > 0$, there exists a constant $C_\nu > 0$ such that*

$$|\phi_g(\mathbf{x})| \leq C_\nu\, e^{-\nu|\mathbf{x}|}, \quad \mathbf{x} \in \mathbb{R}^d, \qquad d = 1, 2, 3. \tag{3.8}$$

3.2. Approximations under the harmonic potential. For any fixed $\beta \geq 0$ in (2.28), we denote the positive ground state of (3.5) with (2.26) as $\phi_g := \phi_g^\beta$ and the corresponding energy and chemical potential as $E_g := E_g^\beta = E(\phi_g^\beta)$ and $\mu_g := \mu_g^\beta = \mu(\phi_g^\beta)$, respectively. When $\beta = 0$, i.e. linear case, the exact ground state ϕ_g^0 can be found as [7, 12, 65–67]

$$E_g^0 = \mu_g^0 = \frac{1}{2}\begin{cases} 1, \\ 1+\gamma_y, \\ 1+\gamma_y+\gamma_z, \end{cases} \qquad \phi_g^0(\mathbf{x}) = \begin{cases} \frac{1}{\pi^{1/4}}\, e^{-x^2/2}, & d=1, \\ \frac{\gamma_y^{1/4}}{\pi^{1/2}}\, e^{-(x^2+\gamma_y y^2)/2}, & d=2, \\ \frac{(\gamma_y\gamma_z)^{1/4}}{\pi^{3/4}}\, e^{-(x^2+\gamma_y y^2+\gamma_z z^2)/2}, & d=3. \end{cases}$$

When $|\beta| = o(1)$ in (2.28), i.e. weak interaction case, the ground state ϕ_g^β can be approximated by $\phi_g^\beta(\mathbf{x}) \approx \phi_g^0(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$, and the corresponding energy E_g^β and chemical potential μ_g^β can be approximated with $C_d = \int_{\mathbb{R}^d} |\phi_g^0(\mathbf{x})|^4\, d\mathbf{x}$ as

$$E_g^\beta \approx E(\phi_g^0) = E_g^0 + \frac{\beta}{2} C_d = E_g^0 + O(\beta), \quad \mu_g^\beta \approx \mu(\phi_g^0) = \mu_g^0 + \beta C_d = \mu_g^0 + O(\beta),$$

where $C_1 = \sqrt{\pi/2}$, $C_2 = \sqrt{\gamma_y}/2\pi$ and $C_3 = \sqrt{\gamma_y\gamma_z}/(2\pi)^{3/2}$.

When $\beta \gg 1$, the ground state ϕ_g^β can be well approximated by the Thomas-Fermi (TF) approximation $\phi_g^\beta \approx \phi_g^{\rm TF}$ [7, 67], i.e. by dropping the diffusion term (e.g. the first term on the right hand side of (3.2)), we obtain

$$\mu_g^{\rm TF}\phi_g^{\rm TF}(\mathbf{x}) = V(\mathbf{x})\phi_g^{\rm TF}(\mathbf{x}) + \beta|\phi_g^{\rm TF}(\mathbf{x})|^2\phi_g^{\rm TF}(\mathbf{x}), \qquad \mathbf{x}\in\mathbb{R}^d, \tag{3.9}$$

with $\mu_g^{\rm TF} \approx \mu_g^\beta$. Solving the above equation, we get

$$\phi_g^\beta(\mathbf{x}) \approx \phi_g^{\rm TF}(\mathbf{x}) = \begin{cases} \sqrt{\left(\mu_g^{\rm TF} - V(\mathbf{x})\right)/\beta}, & V(\mathbf{x}) < \mu_g^{\rm TF}, \\ 0, & \text{otherwise}, \end{cases} \tag{3.10}$$

where $\mu_g^{\rm TF}$ is chosen to satisfy the normalization $\|\phi_g^{\rm TF}\| = 1$, which can be computed as [7, 12, 65–67]

$$\mu_g^\beta \approx \mu_g^{\rm TF} = \begin{cases} \frac{1}{2}\left(\frac{3\beta}{2}\right)^{2/3}, \\ \left(\frac{\beta\gamma_y}{\pi}\right)^{1/2}, \\ \frac{1}{2}\left(\frac{15\beta\gamma_y\gamma_z}{4\pi}\right)^{2/5}, \end{cases} \qquad E_g^\beta \approx E_g^{\rm TF} = \begin{cases} \frac{3}{10}\left(\frac{3\beta}{2}\right)^{2/3}, & d=1, \\ \frac{2}{3}\left(\frac{\beta\gamma_y}{\pi}\right)^{1/2}, & d=2, \\ \frac{5}{14}\left(\frac{15\beta\gamma_y\gamma_z}{4\pi}\right)^{2/5}, & d=3, \end{cases}$$

with $E_g^{\rm TF} := \mu_g^{\rm TF} - E_{\rm int}(\phi_g^{\rm TF})$. For fixed $\gamma_y \geq 1$ and $\gamma_z \geq 1$ in (2.26) and when $\beta \gg 1$ (e.g. $N \gg 1$), from the above TF approximation, we can get the typical lengthes (i.e. $R_x^{\rm TF} = \sqrt{2\mu_g^{\rm TF}}$, $R_y^{\rm TF} = \sqrt{2\mu_g^{\rm TF}/\gamma_y}$ and $R_z^{\rm TF} = \sqrt{2\mu_g^{\rm TF}/\gamma_z}$ of the support of the TF approximation $\phi_g^{\rm TF}$ in x-, y- and z-directions, respectively) – TF radius– of the ground state ϕ_g^β for a BEC as: $R_x^{\rm TF} = O(\beta^{1/(d+2)}) = O(N^{1/(d+2)})$ for $d = 1, 2, 3$, $R_y^{\rm TF} = O(\beta^{1/(d+2)}) = O(N^{1/(d+2)})$ for $d = 2, 3$, and $R_z^{\rm TF} = O(\beta^{1/5}) = O(N^{1/5})$ for $d = 3$. In addition, we also have $E_g^\beta \approx E_g^{\rm TF} = \frac{d+2}{d+4}\mu_g^{\rm TF} \approx \frac{d+2}{d+4}\mu_g^\beta = O(\beta^{2/(d+2)}) = O(N^{2/(d+2)})$, $\|\phi_g^\beta\|_{L^\infty} \approx \phi_g^{\rm TF}(\mathbf{0}) = O(\beta^{-d/2(d+2)}) = O(N^{-d/2(d+2)})$ for $d = 1, 2, 3$. Thus it is easy to see that there is no limit of the ground state ϕ_g^β when $\beta \to \infty$ under the standard physics scaling (2.10) for a BEC. In addition, for computing the ground states and dynamics of a BEC, the bounded computational domain needs to be chosen depending on β such that the truncation error can be negligible!

3.3. Numerical methods. Various numerical methods for computing the ground state ϕ_g in (3.5) have been proposed and studied in the literature [7, 11, 12, 22, 34, 64]. Among them, one of the most efficient and simple methods is the following *gradient flow with discrete normalization* (GFDN) [7, 12]. Choose a time step $\tau := \Delta t > 0$ and denote time steps as $t_n = n\tau$ for $n = 0, 1, \ldots$ At each time interval $[t_n, t_{n+1})$, by applying the steepest decent method to the energy functional $E(\phi)$ without constraint and then projecting the solution back to the unit sphere S at $t = t_{n+1}$ so as to satisfy the constraint (3.3), we have

$$\partial_t\phi = -\frac{1}{2}\frac{\delta E(\phi)}{\delta\phi} = \left[\frac{1}{2}\nabla^2 - V(\mathbf{x}) - \beta\,|\phi|^2\right]\phi, \quad t_n < t < t_{n+1}, \tag{3.11}$$

$$\phi(\mathbf{x}, t_{n+1}) \triangleq \phi(\mathbf{x}, t_{n+1}^+) = \frac{\phi(\mathbf{x}, t_{n+1}^-)}{\|\phi(\cdot, t_{n+1}^-)\|}, \qquad \mathbf{x}\in\mathbb{R}^d, \quad n \geq 0, \tag{3.12}$$

where $\phi := \phi(\mathbf{x}, t)$, $\phi(\mathbf{x}, t_n^{\pm}) = \lim_{t \to t_n^{\pm}} \phi(\mathbf{x}, t)$, and with the initial data

$$\phi(\mathbf{x}, 0) = \phi_0(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d. \tag{3.13}$$

In fact, the gradient flow (3.11) can be obtained from the GPE (2.25) by $t \to -it$, thus the GFDN is known as *imaginary time method* in physics literatures [34, 64].

For the above GFDN, suppose $V(\mathbf{x}) \geq 0$ for $\mathbf{x} \in \mathbb{R}^d$ and $\|\phi_0\|^2 := \int_{\mathbb{R}^d} |\phi_0(\mathbf{x})|^2 \, d\mathbf{x} = 1$, then we have [7, 12]

Theorem 3.5 (Energy diminishing [12]). *For $\beta = 0$, the GFDN (3.11)-(3.13) is energy diminishing for any time step $\tau > 0$ and initial data ϕ_0, i.e.*

$$E(\phi(\cdot, t_{n+1})) \leq E(\phi(\cdot, t_n)) \leq \cdots \leq E(\phi(\cdot, 0)) = E(\phi_0), \quad n = 0, 1, 2, \cdots. \tag{3.14}$$

Let $\tau \to 0$ in (3.11)-(3.13), we can obtain the following *normalized gradient flow* (NGF) [12]

$$\partial_t \phi(\mathbf{x}, t) = \left[\frac{1}{2} \nabla^2 - V(\mathbf{x}) - \beta \, |\phi|^2 + \mu_\phi(t) \right] \phi, \qquad \mathbf{x} \in \mathbb{R}^d, \quad t \geq 0, \tag{3.15}$$

where

$$\mu_\phi(t) = \frac{\mu(\phi(\cdot, t))}{\|\phi(\cdot, t)\|^2} = \frac{1}{\|\phi(\cdot, t)\|^2} \int_{\mathbb{R}^d} \left[\frac{1}{2} |\nabla \phi|^2 + V(\mathbf{x}) |\phi|^2 + \beta |\phi|^4 \right] d\mathbf{x}. \tag{3.16}$$

Theorem 3.6 (Energy diminishing [12]). *The NGF (3.15) with (3.13) is normalization conservative and energy diminishing, i.e.*

$$\|\phi(\cdot, t)\| \equiv \|\phi_0\| = 1, \quad \frac{d}{dt} E(\phi) = -2 \left\| \partial_t \phi(\cdot, t) \right\|^2 \leq 0 \, , \qquad t \geq 0, \tag{3.17}$$

which in turn implies

$$E(\phi(\cdot, t)) \geq E(\phi(\cdot, s)), \qquad 0 \leq t \leq s < \infty. \tag{3.18}$$

With the above two theorems, the positive ground state can be obtained from the GFDN as $\phi_g(\mathbf{x}) = \lim_{t \to \infty} \phi(\mathbf{x}, t)$ provided that ϕ_0 is chosen as a positive function and time step τ is not too big when $\beta \geq 0$ [7, 12]. In addition, the GFDN (3.11)-(3.13) can be discretized by the *backward Euler finite difference* (BEFD) discretization [7, 12]. For simplicity of notation, here we only present the BEFD for the GFDN in 1D truncated on a bounded interval $U = (a, b)$ with homogeneous Dirichlet boundary conditions. Choose a mesh size $h := \Delta x = (b - a)/M > 0$ with M a positive integer, denote grid points as $x_j = a + jh$ for $j = 0, 1, \ldots, M$, and let ϕ_j^n be the numerical approximation of $\phi(x_j, t_n)$. Then a BEFD discretization for the GFDN in 1D reads [7, 12]

$$\frac{\phi_j^{(1)} - \phi_j^n}{\tau} = \frac{\phi_{j+1}^{(1)} - 2\phi_j^{(1)} + \phi_{j-1}^{(1)}}{2h^2} - \left[V(x_j) + \beta \left(\phi_j^n \right)^2 \right] \phi_j^{(1)}, \quad 1 \leq j \leq M - 1,$$

$$\phi_0^{(1)} = \phi_M^{(1)} = 0, \quad \phi_j^0 = \phi_0(x_j), \quad \phi_j^{n+1} = \frac{\phi_j^{(1)}}{\|\phi^{(1)}\|_h}, \quad 0 \leq j \leq M, \quad n \geq 0,$$

where $\|\phi^{(1)}\|_h^2 := h \sum_{j=1}^{M-1} |\phi_j^{(1)}|^2$. This BEFD method is implicit and unconditionally stable, the discretized system can be solved by the Thomas' algorithm, the memory cost is

$O(M)$ and computational cost is $O(M)$ per time step. The ground state can be obtained numerically from the above BEFD when $\max\limits_{0\le j\le M} \frac{|\phi_j^{n+1}-\phi_j^n|}{\tau} \le \varepsilon$ with ε small enough, e.g. 10^{-6}. For extensions to 2D and 3D as well as other numerical methods, we refer [7, 11, 12, 22, 34, 64] and references therein.

4. Dynamics

For studying the dynamics of the GPE (2.25), the initial data is usually chosen as

$$\psi(\mathbf{x},0)=\psi_0(\mathbf{x}), \qquad \mathbf{x}\in\mathbb{R}^d. \tag{4.1}$$

The GPE (2.25) is a dispersive PDE and it is *time reversible or symmetric*, i.e. it is unchanged under the change of variable in time as $t\to -t$ and taken conjugate in the equation. Another important property is *time transverse or gauge invariant*, i.e. if $V\to V+\alpha$ with α a given real constant, then the solution $\psi\to \psi e^{-i\alpha t}$ which immediately implies that the density $\rho=|\psi|^2$ is unchanged. It conserves the normalization (or mass) and energy (or Hamiltonian), i.e. $N(\psi(\cdot,t))\equiv N(\psi_0)$ and $E(\psi(\cdot,t))\equiv E(\psi_0)$ for $t\ge 0$.

4.1. Well-posedness and dynamical properties. For studying well-posedness of the GPE (2.25), we introduce the functional spaces

$$L_V(\mathbb{R}^d)=\left\{\phi\big|\int_{\mathbb{R}^d} V(\mathbf{x})|\phi(\mathbf{x})|^2 d\mathbf{x}<\infty\right\},\quad X:=X(\mathbb{R}^d)=H^1(\mathbb{R}^d)\cap L_V(\mathbb{R}^d).$$

Theorem 4.1 (Well-posedness [7]). *Suppose the trapping potential is nonnegative and at most quadratic growth in far field, i.e., $V(\mathbf{x})\in C^\infty(\mathbb{R}^d)$ and $D^{\mathbf{k}}V(\mathbf{x})\in L^\infty(\mathbb{R}^d)$ for all $\mathbf{k}\in\mathbb{N}_0^d$ with $|\mathbf{k}|\ge 2$, then we have*

(i) *For any initial data $\psi_0\in X(\mathbb{R}^d)$, there exists a time $T_{\max}\in(0,+\infty]$ such that the Cauchy problem of the GPE (2.25) with (4.1) has a unique maximal solution $\psi\in C([0,T_{\max}),X)$. It is maximal in the sense that if $T_{\max}<\infty$, then $\|\psi(\cdot,t)\|_X\to\infty$ when $t\to T_{\max}^-$.*

(ii) *As long as the solution $\psi(\mathbf{x},t)$ remains in the energy space X, the L^2-norm $\|\psi(\cdot,t)\|_2$ and energy $E(\psi(\cdot,t))$ are conserved for $t\in[0,T_{\max})$.*

(iii) *The solution of the Cauchy problem is global in time, i.e., $T_{\max}=\infty$, if $d=1$ or $d=2$ with $\beta>C_b/\|\psi_0\|_2^2$ or $d=3$ with $\beta\ge 0$.*

Theorem 4.2 (Finite time blow-up [7]). *In 2D and 3D, assume $V(\mathbf{x})$ is at most quadratic growth in far field and satisfies $V(\mathbf{x})d+\mathbf{x}\cdot\nabla V(\mathbf{x})\ge 0$ for $\mathbf{x}\in\mathbb{R}^d$ ($d=2,3$). When $\beta<0$, for any initial data $\psi_0(\mathbf{x})\in X$ with finite variance $\int_{\mathbb{R}^d}|\mathbf{x}|^2|\psi_0|^2\,d\mathbf{x}<\infty$, the Cauchy problem of the GPE (2.25) with (4.1) will blow-up at finite time, i.e. $T_{\max}<\infty$, if one of the following holds:*

(i) $E(\psi_0)<0$;

(ii) $E(\psi_0)=0$ *and* $\mathrm{Im}\left(\int_{\mathbb{R}^d}\overline{\psi}_0(\mathbf{x})\,(\mathbf{x}\cdot\nabla\psi_0(\mathbf{x}))\,d\mathbf{x}\right)<0$;

(iii) $E(\psi_0)>0$ *and* $\mathrm{Im}\left(\int_{\mathbb{R}^d}\overline{\psi}_0(\mathbf{x})\,(\mathbf{x}\cdot\nabla\psi_0(\mathbf{x}))\,d\mathbf{x}\right)<-\sqrt{E(\psi_0)d}\,\|\mathbf{x}\psi_0\|_{L^2}$.

If there is no external potential in the GPE (2.25), i.e. $V(\mathbf{x}) \equiv 0$, then the momentum and angular momentum are also conserved [4, 7, 70]. The GPE (2.25) admits the plane wave solution as $\psi(\mathbf{x}, t) = Ae^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$, where the time frequency ω, amplitude A and spatial wave number $\mathbf{k}$ satisfy the following *dispersion relation* [4, 7, 70]: $\omega = \frac{|\mathbf{k}|^2}{2} + \beta|A|^2$. In 1D, i.e. $d = 1$, when $\beta < 0$, it admits the well-known bright soliton solution as [4, 70]

$$\psi_B(x, t) = \frac{A}{\sqrt{-\beta}}\text{sech}(A(x - vt - x_0))e^{i(vx-\frac{1}{2}(v^2-A^2)t+\theta_0)}, \qquad x \in \mathbb{R}, \quad t \geq 0, \tag{4.2}$$

where $\frac{A}{\sqrt{-\beta}}$ is the amplitude of the soliton with A a positive real constant, v is the velocity of the soliton, x_0 and θ_0 are the initial shifts in space and phase, respectively. Since the soliton solution is exponentially decaying for $|x| \to +\infty$, then the mass and energy are well defined and given by: $N(\psi_B) = -\frac{2A}{\beta}$ and $E(\psi_B) = \frac{Av^2}{-\beta} + \frac{A^3}{-3\beta}$. When $\beta > 0$, it admits dark solitons [67, 70].

Let $\psi := \psi(\mathbf{x}, t)$ be the solution of the GPE (2.25) with the harmonic potential (2.26) and initial data (4.1) satisfying $\|\psi_0\| = 1$, define the center-of-mass $\mathbf{x}_c(t) = \int_{\mathbb{R}^d} \mathbf{x}|\psi(\mathbf{x}, t)|^2\, d\mathbf{x}$, square of the condensate width $\delta_\alpha(t) = \int_{\mathbb{R}^d} \alpha^2|\psi(\mathbf{x}, t)|^2 d\mathbf{x}$ with $\alpha = x$, y or z, and angular momentum expectation $\langle L_z\rangle(t) = \int_{\mathbb{R}^d} \overline{\psi(\mathbf{x}, t)}L_z\psi(\mathbf{x}, t)\, d\mathbf{x}$ with $L_z = -i\,(x\partial_y - y\partial_x)$ when $d = 2, 3$. Then we have [7, 13]

Lemma 4.3 (Angular momentum expectation [7, 13]). *For any initial data $\psi_0(\mathbf{x})$ in (4.1), when $\gamma_y = 1$ in (2.26), i.e. the trapping potential is radially/cylindrically symmetric in 2D/3D, then the angular momentum expectation is conserved, i.e.*

$$\langle L_z\rangle(t) \equiv \langle L_z\rangle(0) = \int_{\mathbb{R}^d} \overline{\psi_0(\mathbf{x})}L_z\psi_0(\mathbf{x})\, d\mathbf{x}, \qquad t \geq 0. \tag{4.3}$$

Lemma 4.4 (Condensate width [7, 13]). *For any initial data $\psi_0(\mathbf{x})$ in (4.1), in 1D without interaction, i.e. $d = 1$ and $\beta = 0$ in (2.25), we have*

$$\delta_x(t) = E(\psi_0) + \Big(\delta_x^{(0)} - E(\psi_0)\Big)\cos(2t) + \delta_x^{(1)}\sin(2t), \quad t \geq 0; \tag{4.4}$$

and in 2D with a radially symmetric trap, i.e. $d = 2$ and $\gamma_y = 1$ in (2.26), we have

$$\delta_r(t) = E(\psi_0) + \Big(\delta_r^{(0)} - E(\psi_0)\Big)\cos(2t) + \delta_r^{(1)}\sin(2t), \quad t \geq 0, \tag{4.5}$$

where $\delta_r(t) = \delta_x(t) + \delta_y(t)$, $\delta_r^{(0)} := \delta_x^{(0)} + \delta_y^{(0)}$, and $\delta_r^{(1)} := \delta_x^{(1)} + \delta_y^{(1)}$ with $\delta_\alpha^{(0)} = \int_{\mathbb{R}^d} \alpha^2|\psi_0(\mathbf{x})|^2 d\mathbf{x}$ and $\delta_\alpha^{(1)} = 2\int_{\mathbb{R}^d} \alpha\,\text{Im}\left(\overline{\psi}_0\partial_\alpha\psi_0\right)\, d\mathbf{x}$ for $\alpha = x$ or y. Thus δ_x in 1D and δ_r in 2D are periodic functions with frequency doubling the trapping frequency.

Lemma 4.5 (Center-of-mass [7, 13, 19]). *For any initial data $\psi_0(\mathbf{x})$ in (4.1), the dynamics of the center-of-mass satisfies the following second-order ODE*

$$\ddot{\mathbf{x}}_c(t) + \Lambda\,\mathbf{x}_c(t) = 0, \quad t \geq 0, \tag{4.6}$$

with the following initial data

$$\mathbf{x}_c(0) = \mathbf{x}_c^{(0)} = \int_{\mathbb{R}^d} \mathbf{x}|\psi_0(\mathbf{x})|^2 d\mathbf{x}, \quad \dot{\mathbf{x}}_c(0) = \mathbf{x}_c^{(1)} = \int_{\mathbb{R}^d} \text{Im}(\overline{\psi}_0\nabla\psi_0)\, d\mathbf{x},$$

where Λ is a $d \times d$ diagonal matrix as $\Lambda = 1$ when $d = 1$, $\Lambda = \text{diag}(1, \gamma_y^2)$ when $d = 2$, and $\Lambda = \text{diag}(1, \gamma_y^2, \gamma_z^2)$ when $d = 3$. This implies that each component of $\mathbf{x}_c$ is a periodic function whose frequency is the same as the trapping frequency in that direction.

Lemma 4.6 (Exact solution [7, 13]). *If the initial data $\psi_0(\mathbf{x})$ in (4.1) is chosen as*

$$\psi_0(\mathbf{x}) = \phi_e(\mathbf{x}-\mathbf{x}_0)\, e^{i(\mathbf{w}_0\cdot\mathbf{x}+g_0)}, \qquad \mathbf{x}\in\mathbb{R}^d, \tag{4.7}$$

where $\mathbf{x}_0, \mathbf{w}_0 \in \mathbb{R}^d$ and $g_0 \in \mathbb{R}$ are given constants, and (μ_e, ϕ_e) is a solution of the nonlinear eigenvalue problem (3.2) with the constraint (3.3), then the GPE (2.25) with (2.26) and (4.7) admits the following unique exact solution

$$\psi(\mathbf{x},t) = \phi_e(\mathbf{x}-\mathbf{x}_c(t))\, e^{-i\mu_e t}\, e^{i(\mathbf{w}(t)\cdot\mathbf{x}+g(t))}, \qquad \mathbf{x}\in\mathbb{R}^d, \quad t\ge 0, \tag{4.8}$$

where $\mathbf{x}_c(t)$ satisfies the second-order ODE (4.6) with the initial condition $\mathbf{x}_c(0)=\mathbf{x}_0$ and $\dot{\mathbf{x}}_c(0)=\mathbf{w}_0$, and $\mathbf{w}(t)$ and $g(t)$ satisfy the following ODEs

$$\dot{\mathbf{w}}(t) = -\Lambda\,\mathbf{x}_c(t), \qquad \dot{g}(t) = V(\mathbf{x}_c(t)) = \frac{1}{2}\mathbf{x}_c(t)\cdot(\Lambda\,\mathbf{x}_c(t)), \qquad t>0, \tag{4.9}$$

with initial data $\mathbf{w}(0)=\mathbf{w}_0$ and $g(0)=g_0$.

4.2. Numerical methods. Various numerical methods have been proposed and studied in the literature [4, 7, 14, 20, 34, 64] for computing the dynamics of the GPE (2.25) with (4.1). Among them, one of the most efficient and accurate as well as simple methods is the following *time-splitting sine pseudospectral* (TSSP) method [4, 7, 14]. For simplicity of notation, here we only present the TSSP method for the GPE (2.25) in 1D truncated on a bounded interval $U=(a,b)$ with homogeneous Dirichlet boundary conditions. Let ψ_j^n be the numerical approximation of $\psi(x_j,t_n)$ and ψ^n be the solution vector at time $t=t_n=n\tau$ with components $\{\psi_j^n\}_{j=0}^M$, then a second-order TSSP method for the GPE (2.25) in 1D reads [4, 7, 14]

$$\psi_j^{(1)} = \frac{2}{M}\sum_{l=1}^{M-1} e^{-i\tau\mu_l^2/4}\,\widetilde{(\psi^n)}_l\,\sin(\mu_l(x_j-a)), \quad \psi_j^{(2)} = e^{-i\tau(V(x_j)+\beta|\psi_j^{(1)}|^2)}\,\psi_j^{(1)},$$

$$\psi_j^{n+1} = \frac{2}{M}\sum_{l=1}^{M-1} e^{-i\tau\mu_l^2/4}\,\widetilde{(\psi^{(2)})}_l\,\sin(\mu_l(x_j-a)), \qquad 0\le j\le M,$$

where $\mu_l = l\pi/(b-a)$ for $1\le l\le M-1$ and $\widetilde{(\psi^n)}_l$ and $\widetilde{(\psi^{(2)})}_l$ are the discrete sine transform (DST) coefficients of ψ^n and $\psi^{(2)}$, respectively. This TSSP method for the GPE (2.25) is explicit, unconditionally stable, second-order accurate in time and spectral-order accurate in space [4, 7, 14]. It is time reversible or symmetric, time transverse invariant, conserves the mass at the discetized level and has the same dispersive relation as the GPE when $V(\mathbf{x})\equiv 0$. The memory cost is $O(M)$ and computational cost is $O(M\ln M)$ per time step. For extensions to 2D/3D and other numerical methods, we refer to [4, 7, 14, 20, 34, 64] and references therein.

4.3. Bogoliubov excitation of ground state. An important class of time-dependent solutions of the GPE (2.25) is given by the small-amplitude oscillations, where the changes in space and time of the wave function (or order parameter) with respect to the stationary states, especially ground states, are small. In many cases these solutions emphasize the collective behavior exhibited by the interacting Bose gases and can be interpreted in terms of the elementary excitations of the system. For describing the dynamics of a BEC, it is natural

to consider the linearized behavior of small perturbations around its ground state ϕ_g with chemical potential μ_g and take the ansatz [38, 43, 45, 67]

$$\psi(\mathbf{x},t) = e^{-i\mu_g t}\left[\phi_g(\mathbf{x}) + u(\mathbf{x})e^{-i\omega t} - \overline{v(\mathbf{x})}e^{i\omega t}\right], \qquad \mathbf{x}\in\mathbb{R}^d, \quad t>0, \tag{4.10}$$

where the Bogoliubov amplitudes $u(\mathbf{x})$ and $v(\mathbf{x})$ are treated as small and $\omega \in \mathbb{C}$ to be determined. Substituting (4.10) into (2.25) and collecting first-order terms proportional to $e^{\pm i\omega t}$, we obtain the Bogoliubov equations – linear eigenvalue problem for (ω, u, v)— as [38, 43, 45, 67]

$$\begin{aligned} \omega\, u(\mathbf{x}) &= \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + 2\beta|\phi_g(\mathbf{x})|^2 - \mu_g\right] u(\mathbf{x}) - \phi_g^2\, v(\mathbf{x}), \quad \mathbf{x}\in\mathbb{R}^d, \\ -\omega\, v(\mathbf{x}) &= \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + 2\beta|\phi_g(\mathbf{x})|^2 - \mu_g\right] v(\mathbf{x}) - \overline{\phi}_g^2\, u(\mathbf{x}). \end{aligned} \tag{4.11}$$

In many ways, the above Bogoliubov equations are analogous to a nonrelativistic version of the Dirac equation, with u and v as the particle and hole amplitudes, including the $(+,-)$ metric seen in the minus sign on the left hand side of the second equation compared to the first equation in (4.11) [38, 43, 45, 67]. In addition, a detailed analysis shows that physically relevant Bogoliubov eigenfunctions must satisfy the following *positive* normalization condition [38, 43, 45, 67]:

$$\|u\|^2 - \|v\|^2 := \int_{\mathbb{R}^d}\left[|u(\mathbf{x})|^2 - |v(\mathbf{x})|^2\right] d\mathbf{x} = 1. \tag{4.12}$$

For solutions of the Bogoliubov equations, especially no external trapping potential in (2.25), we refer to [38, 43, 45, 67] and references therein.

4.4. Semiclassical scaling and limits. In the strongly repulsive interaction regime, i.e. $\beta \gg 1$ in the GPE (2.25) with (2.26), another scaling (under the normalization (2.27) with ψ being replaced by ψ^ε) – semiclassical scaling – is also very useful in practice, especially in numerical computation. By choosing $\mathbf{x} \to \mathbf{x}\varepsilon^{-1/2}$ and $\psi = \varepsilon^{d/4}\psi^\varepsilon$ with $0<\varepsilon = 1/\beta^{2/(2+d)} < 1$ ($\Leftrightarrow t = \frac{1}{\omega_x}$, $x_s = \sqrt{\hbar/m\varepsilon\omega_x}$ and $E_s = \hbar\omega_x/\varepsilon$ in (2.10) for the GPE (2.8) when $d=3$), we obtain [7, 14]

$$i\varepsilon\,\partial_t\psi^\varepsilon(\mathbf{x},t) = \left[-\frac{\varepsilon^2}{2}\nabla^2 + V(\mathbf{x}) + |\psi^\varepsilon(\mathbf{x},t)|^2\right]\psi^\varepsilon(\mathbf{x},t), \quad \mathbf{x}\in\mathbb{R}^d, \ t>0. \tag{4.13}$$

This GPE conserves the following energy

$$E^\varepsilon(\psi^\varepsilon(\cdot,t)) = \int_{\mathbb{R}^d}\left[\frac{\varepsilon^2}{2}|\nabla\psi^\varepsilon|^2 + V(\mathbf{x})|\psi^\varepsilon|^2 + \frac{1}{2}|\psi^\varepsilon|^4\right] d\mathbf{x} \equiv E^\varepsilon(\psi^\varepsilon(\cdot,0)), \quad t\ge 0.$$

Similarly, the nonlinear eigenvalue problem (3.2) (under the normalization (3.3) with $\phi = \phi^\varepsilon$) reads

$$\mu^\varepsilon\phi^\varepsilon(\mathbf{x}) = \left[-\frac{\varepsilon^2}{2}\nabla^2 + V(\mathbf{x}) + |\phi^\varepsilon(\mathbf{x})|^2\right]\phi^\varepsilon(\mathbf{x}), \quad \mathbf{x}\in\mathbb{R}^d, \tag{4.14}$$

where the eigenvalue (or chemical potential) μ^ε can be computed from its corresponding eigenfunction ϕ^ε by $\mu^\varepsilon = \mu^\varepsilon(\phi^\varepsilon) = E^\varepsilon(\phi^\varepsilon) + E^\varepsilon_{\rm int}(\phi^\varepsilon)$ with $E^\varepsilon_{\rm int}(\phi^\varepsilon) = \frac{1}{2}\int_{\mathbb{R}^d}|\phi^\varepsilon|^4\, d\mathbf{x}$.

The constrained minimization problem for ground state collapses to: Find $\phi_g^\varepsilon \in S$ such that

$$E_g^\varepsilon := E^\varepsilon(\phi_g^\varepsilon) = \min_{\phi^\varepsilon \in S} E^\varepsilon(\phi^\varepsilon), \qquad \text{with } \mu_g^\varepsilon := \mu^\varepsilon(\phi_g^\varepsilon) = E^\varepsilon(\phi_g^\varepsilon) + E_{\rm int}^\varepsilon(\phi_g^\varepsilon). \quad (4.15)$$

Similarly to section 3.2, we can get the TF approximation to the ground state when $0 < \varepsilon \ll 1$:

$$\phi_g^\varepsilon(\mathbf{x}) \approx \phi_g^{\rm TF}(\mathbf{x}) = \begin{cases} \sqrt{\mu_g^{\rm TF} - V(\mathbf{x})}, & V(\mathbf{x}) < \mu_g^{\rm TF}, \\ 0, & \text{otherwise}, \end{cases} \quad (4.16)$$

where

$$\mu_g^\varepsilon \approx \mu_g^{\rm TF} = \begin{cases} \frac{1}{2}\left(\frac{3}{2}\right)^{2/3}, \\ \left(\frac{\gamma_y}{\pi}\right)^{1/2}, \\ \frac{1}{2}\left(\frac{15\gamma_y\gamma_z}{4\pi}\right)^{2/5}, \end{cases} \qquad E_g^\varepsilon \approx E_g^{\rm TF} = \begin{cases} \frac{3}{10}\left(\frac{3}{2}\right)^{2/3}, & d=1, \\ \frac{2}{3}\left(\frac{\gamma_y}{\pi}\right)^{1/2}, & d=2, \\ \frac{5}{14}\left(\frac{15\gamma_y\gamma_z}{4\pi}\right)^{2/5}, & d=3. \end{cases}$$

From this TF approximation, for fixed $\gamma_y \geq 1$ and $\gamma_z \geq 1$ in (2.26) and when $0 < \varepsilon \ll 1$, we have $E_g^\varepsilon \approx E_g^{\rm TF} = \frac{d+2}{d+4}\mu_g^{\rm TF} \approx \frac{d+2}{d+4}\mu_g^\varepsilon = O(1)$, $\|\phi_g^\varepsilon\|_{L^\infty} \approx \phi_g^{\rm TF}(\mathbf{0}) = O(1)$, and the TF radius $R_x^{\rm TF} = \sqrt{2\mu_g^{\rm TF}} = O(1)$, $R_y^{\rm TF} = \sqrt{2\mu_g^{\rm TF}}/\gamma_y = O(1)$ and $R_z^{\rm TF} = \sqrt{2\mu_g^{\rm TF}}/\gamma_z = O(1)$ for $d = 1, 2, 3$. In addition, the ground state $\phi_g^\varepsilon(\mathbf{x})$ converges to $\phi_g^{\rm TF}(\mathbf{x})$ uniformly when $\varepsilon \to 0^+$. Furthermore, for computing numerically the ground states and dynamics of a BEC, the bounded computational domain can be chosen independent of ε [7, 14].

Taking the WKB ansatz $\psi^\varepsilon(\mathbf{x}, t) = \sqrt{\rho^\varepsilon(\mathbf{x}, t)}\, e^{iS^\varepsilon(\mathbf{x},t)/\varepsilon}$ with $\rho^\varepsilon = |\psi^\varepsilon|^2$ and S^ε the density and phase of the wave function, respectively, inserting it into the GPE (4.13) and separating real and imaginary parts, we obtain the transport and Hamilton-Jacobi equations for density and phase, respectively [7, 32, 44]

$$\begin{aligned} &\partial_t \rho^\varepsilon + \operatorname{div}(\rho^\varepsilon\, \nabla S^\varepsilon) = 0, && \mathbf{x} \in \mathbb{R}^d, \quad t > 0, \\ &\partial_t S^\varepsilon + \frac{1}{2}|\nabla S^\varepsilon|^2 + \rho^\varepsilon + V(\mathbf{x}) = \frac{\varepsilon^2}{2}\,\frac{1}{\sqrt{\rho^\varepsilon}}\,\Delta\sqrt{\rho^\varepsilon}. \end{aligned} \quad (4.17)$$

Furthermore, defining the quantum velocity $\mathbf{u}^\varepsilon = \nabla S^\varepsilon$ and current $\mathbf{J}^\varepsilon = \rho^\varepsilon\, \mathbf{u}^\varepsilon$, we get from (4.17) the Euler system with a third-order dispersion correction term – quantum hydrodynamics (QHD) – as [7, 32, 44]

$$\begin{aligned} &\partial_t \rho^\varepsilon + \operatorname{div} \mathbf{J}^\varepsilon = 0, && \mathbf{x} \in \mathbb{R}^d, \quad t > 0, \\ &\partial_t \mathbf{J}^\varepsilon + \operatorname{div}\left(\frac{\mathbf{J}^\varepsilon \otimes \mathbf{J}^\varepsilon}{\rho^\varepsilon}\right) + \rho^\varepsilon\, \nabla V(\mathbf{x}) + \nabla P(\rho^\varepsilon) = \frac{\varepsilon^2}{4}\nabla\left(\rho^\varepsilon \nabla^2 \ln \rho^\varepsilon\right), \end{aligned} \quad (4.18)$$

where the pressure is defined as $P(\rho^\varepsilon) = (\rho^\varepsilon)^2/2$. Letting $\varepsilon \to 0^+$ in (4.18), formally we get the Euler system [7, 32, 44]

$$\begin{aligned} &\partial_t \rho^0 + \operatorname{div} \mathbf{J}^0 = 0, && \mathbf{x} \in \mathbb{R}^d, \quad t > 0, \\ &\partial_t \mathbf{J}^0 + \operatorname{div}\left(\frac{\mathbf{J}^0 \otimes \mathbf{J}^0}{\rho^0}\right) + \rho^0\, \nabla V(\mathbf{x}) + \nabla P(\rho^0) = 0. \end{aligned} \quad (4.19)$$

For mathematical justification of the passage from the GPE (4.13) to the Euler system (4.19), we refer to [7, 32, 44] and references therein.

5. Extensions

In this section, we will present briefly mathematical models and theories as well as numerical methods for rotating BEC based on the GPE with an angular momentum rotation term, dipolar BEC based on the GPE with a long-range anistropic dipole-dipole interaction (DDI) and spin-orbit-coupled BEC based on coupled GPEs with an internal atomic Josephon junction (JJ) and an spin-orbit coupling term.

5.1. For rotating BEC. At temperatures T much smaller than the critical temperature T_c, following the mean field theory [1, 2, 7, 31, 43, 57, 62, 69], a BEC in the rotational frame is well described by the macroscopic wave function $\psi := \psi(\mathbf{x}, t)$, whose evolution is governed by the GPE with an angular momentum rotation term

$$i\hbar\partial_t\psi = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) - \tilde{\Omega}L_z + Ng|\psi|^2\right]\psi, \qquad \mathbf{x}\in\mathbb{R}^3, \quad t>0, \tag{5.1}$$

where $\tilde{\Omega}$ is the angular velocity, L_z is the z-component angular momentum operator defined as $L_z = -i\hbar\,(x\partial_y - y\partial_x)$ and ψ satisfies the normalization condition (2.5).

Under the harmonic potential (2.9), similarly to the nondimensionalization in section 2.2 and dimension reduction in 2.3 from 3D to 2D when $\omega_z \gg \max\{\omega_x, \omega_y\}$ for a disk-shaped condensate [2, 7, 13, 23], we can obtain the following dimensionless GPE with an angular momentum rotation term in d-dimensions ($d = 2, 3$):

$$i\,\partial_t\psi = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) - \Omega L_z + \beta|\psi|^2\right]\psi, \quad \mathbf{x}\in\mathbb{R}^d, \quad t>0, \tag{5.2}$$

where $\Omega = \tilde{\Omega}/\omega_x$, $\beta = \kappa$ and $\kappa\sqrt{\gamma_z/2\pi}$ when $d = 3$ and 2, respectively, the dimensionless harmonic potential is given in (2.26) for $d = 3, 2$, and the dimensionless angular momentum rotation term is given as $L_z = -i\,(x\partial_y - y\partial_x)$. The GPE (5.2) conserves the normalization (2.5) and energy per particle

$$E(\psi(\cdot,t)) = \int_{\mathbb{R}^d}\left[\frac{1}{2}\left|\nabla\psi\right|^2 + V(\mathbf{x})|\psi|^2 - \Omega\overline{\psi}\,L_z\psi + \frac{\beta}{2}\left|\psi\right|^4\right]d\mathbf{x} \equiv E(\psi(\cdot,0)),\ t\ge 0.$$

The ground state can be defined the same as (3.5) with the above energy functional. For the existence and uniqueness as well as nonexistence, we have [2, 7, 23, 69]

Theorem 5.1 (Existence and uniqueness [2, 7, 23, 69]). *Suppose that $V(\mathbf{x})$ is taken as the harmonic potential in* (2.26)*, then we have*

i) *There exists a ground state of the rotating BEC (5.2) when $|\Omega| < 1$ and $\beta \ge 0$ in 3D or $\beta > -C_b$ in 2D.*

ii) *For any $\beta \ge 0$, there exists a critical rotation velocity $0 < \Omega_c^\beta \le 1$ – first critical rotation speed – depending on β such that: when $\Omega_c^\beta < |\Omega| < 1$, quantized vortices will appear in the ground state ϕ_g.*

iii) *In 2D with $\gamma_y = 1$ (radially symmetric $V(\mathbf{x})$), there exists $\beta_0 > 0$ such that when $\beta \ge \beta_0$, for $|\Omega| < \Omega_{c_1}^\beta$ ($\Omega_{c_1}^\beta$ depends on β), the ground state can be chosen as positive $|\phi_g|$, and $\phi_g(\mathbf{x}) = e^{i\theta_0}|\phi_g(\mathbf{x})|$ for some constant $\theta_0 \in \mathbb{R}$, and the positive ground state ϕ_g is unique.*

iv) *There exists no ground state of the rotating BEC (5.2) if one of the following holds: (a)* $\beta < 0$ *in 3D or* $\beta < -C_b$ *in 2D; (b)* $|\Omega| > 1$.

Remark 5.2. From the various numerical results, for radially symmetric $V(\mathbf{x})$ in 2D (or cylindrically symmetric in 3D) and any fixed $\beta \geq 0$, the *first critical rotation speed* $0 < \Omega_c^\beta \leq 1$ depends on β and: when $|\Omega| < \Omega_c^\beta$, the ground state can be chosen as nonnegative $|\phi_g|$, and $\phi_g(\mathbf{x}) = e^{i\theta_0}|\phi_g(\mathbf{x})|$ for some constant $\theta_0 \in \mathbb{R}$, and the nonnegative ground state ϕ_g is unique; when $\Omega_c^\beta < |\Omega| < 1$, quantized vortices will appear in the ground state ϕ_g; and when $\Omega_c^\beta = |\Omega|$, there exist at least two different ground states – one without quantized vortices and one with quantized vortices. We remark here that a rigorous mathematical justification is still missing.

For more results on the ground state of the rotating BEC (5.2) and efficient and accurate numerical methods for simulation, such as BEFD [7, 23] or BEFP [11], we refer to [2, 7, 9, 23, 43, 69] and references therein. Similarly, for the well-posedness of the Cauchy problem of (5.2) with the initial data (4.1) and its dynamical properties as well as efficient and accurate numerical methods, such as TSADI [21] or TSGLFHP [17], we refer to [4, 7, 43, 69] and references therein. Here we present a different formulation of the GPE (5.2) under the *rotating Lagrangian coordinates* so that the angular momentum rotation term will be removed [19].

For any time $t \geq 0$, let $A(t)$ be an orthogonal rotational matrix defined as

$$A(t) = \begin{pmatrix} \cos(\Omega t) & \sin(\Omega t) \\ -\sin(\Omega t) & \cos(\Omega t) \end{pmatrix}, \ d = 2, \ A(t) = \begin{pmatrix} \cos(\Omega t) & \sin(\Omega t) & 0 \\ -\sin(\Omega t) & \cos(\Omega t) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \ d = 3.$$

It is easy to verify that $A^{-1}(t) = A^T(t)$ for any $t \geq 0$ and $A(0) = I$ with I the identity matrix. For any $t \geq 0$, we introduce the *rotating Lagrangian coordinates* $\widetilde{\mathbf{x}}$ as [19]

$$\widetilde{\mathbf{x}} = A^{-1}(t)\mathbf{x} = A^T(t)\mathbf{x} \quad \Leftrightarrow \quad \mathbf{x} = A(t)\widetilde{\mathbf{x}}, \qquad \mathbf{x} \in \mathbb{R}^d, \tag{5.3}$$

and denote the wave function in the new coordinates as $\varphi := \varphi(\widetilde{\mathbf{x}}, t)$

$$\varphi(\widetilde{\mathbf{x}}, t) := \psi(\mathbf{x}, t) = \psi\left(A(t)\widetilde{\mathbf{x}}, t\right), \qquad \mathbf{x} \in \mathbb{R}^d, \quad t \geq 0. \tag{5.4}$$

Here, we refer the Cartesian coordinates $\mathbf{x}$ as the *Eulerian coordinates*. Plugging (5.3) and (5.4) into (5.2), we obtain the GPE

$$i\partial_t \varphi(\widetilde{\mathbf{x}}, t) = \left[-\frac{1}{2}\nabla^2 + W(\widetilde{\mathbf{x}}, t) + \beta|\varphi(\widetilde{\mathbf{x}}, t)|^2\right]\varphi(\widetilde{\mathbf{x}}, t), \quad \widetilde{\mathbf{x}} \in \mathbb{R}^d, \quad t > 0, \tag{5.5}$$

where $W(\widetilde{\mathbf{x}}, t) = V(A(t)\widetilde{\mathbf{x}})$ for $\widetilde{\mathbf{x}} \in \mathbb{R}^d$ and $t > 0$, which is time-independent, i.e. $W(\widetilde{\mathbf{x}}, t) = V(\widetilde{\mathbf{x}})$ if the harmonic potential (2.26) is radially/cylindically symmetric in 2D/3D, i.e. $\gamma_y = 1$. In addition, the initial data for the GPE (5.5) from (4.1) is

$$\varphi(\widetilde{\mathbf{x}}, 0) = \psi(\mathbf{x}, 0) = \psi_0(\mathbf{x}) := \varphi_0(\mathbf{x}) = \varphi_0(\widetilde{\mathbf{x}}), \qquad \widetilde{\mathbf{x}} = \mathbf{x} \in \mathbb{R}^d. \tag{5.6}$$

Based on the above new formulation, the results and numerical methods developed for non-rotating BEC, such as TSSP [4, 7, 14, 17, 20], can be directly applied for analyzing and simulating the dynamics of rotating BEC.

5.2. For dipolar BEC. At temperature T much smaller than the critical temperature T_c, a dipolar BEC is well described by the macroscopic wave function $\psi := \psi(\mathbf{x}, t)$ whose evolution is governed by the following 3D GPE [6, 7, 10, 24, 55, 71]

$$i\hbar\partial_t\psi = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) + Ng|\psi|^2 + NC_{dd}\left(V_{\rm dip} * |\psi|^2\right)\right]\psi, \quad \mathbf{x} \in \mathbb{R}^3,\ t > 0,$$

where $C_{dd} = \mu_0\mu_{\rm dip}^2/3$ with μ_0 the vacuum magnetic permeability and $\mu_{\rm dip}$ the permanent magnetic dipole moment, ψ satisfies the normalization condition (2.5), and the long-range and anisotropic DDI between two dipoles with the same dipole moment or orientation $\mathbf{n} = (n_1, n_2, n_3)^T \in \mathbb{R}^3$ (which is a given unit vector satisfying $|\mathbf{n}| = \sqrt{n_1^2 + n_2^2 + n_3^3} = 1$) is given by

$$V_{\rm dip}(\mathbf{x}) = \frac{3}{4\pi}\,\frac{1 - 3(\mathbf{x}\cdot\mathbf{n})^2/|\mathbf{x}|^2}{|\mathbf{x}|^3} = \frac{3}{4\pi}\,\frac{1 - 3\cos^2(\theta)}{|\mathbf{x}|^3}, \qquad \mathbf{x} \in \mathbb{R}^3, \tag{5.7}$$

where θ is the angle between the dipole axis $\mathbf{n}$ and the vector $\mathbf{x}$. We remark here that it is still an open problem to derive the above GPE from the N-body linear Schrödinger equation (2.3) with $V_{\rm int}$ in (2.2) is taken as $V_{\rm dip}$.

Again, under the harmonic potential (2.9), similarly to the nondimensionalization in section 2.2 and dimension reduction in 2.3 from 3D to 2D when $\omega_z \gg \max\{\omega_x, \omega_y\}$ for a disk-shaped condensate and to 1D when $\omega_z = \omega_y \gg \omega_x$ for a cigar-shaped condensate [6, 7, 30], by using the decomposition of contact and long-range (or repulsive and attractive) parts of the DDI (5.7) [10, 30]

$$U_{\rm dip}(\mathbf{x}) = \frac{3}{4\pi|\mathbf{x}|^3}\left(1 - \frac{3(\mathbf{x}\cdot\mathbf{n})^2}{|\mathbf{x}|^2}\right) = -\delta(\mathbf{x}) - 3\partial_{\mathbf{nn}}\left(\frac{1}{4\pi|\mathbf{x}|}\right), \quad \mathbf{x} \in \mathbb{R}^3, \tag{5.8}$$

where the differential operators $\partial_{\mathbf{n}} = \mathbf{n}\cdot\nabla$ and $\partial_{\mathbf{nn}} = \partial_{\mathbf{n}}\partial_{\mathbf{n}}$, we can obtain the following dimensionless GPE with a DDI in d-dimensions ($d = 1, 2, 3$):

$$\begin{aligned} &i\partial_t\psi(\mathbf{x}, t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + \beta|\psi(\mathbf{x}, t)|^2 + \eta\varphi(\mathbf{x}, t)\right]\psi(\mathbf{x}, t), \\ &\varphi(\mathbf{x}, t) = L_{\mathbf{n}}u(\mathbf{x}, t), \qquad u(\mathbf{x}, t) = G * |\psi|^2, \qquad\qquad \mathbf{x} \in \mathbb{R}^d, \quad t \geq 0, \end{aligned} \tag{5.9}$$

where

$$\beta = \begin{cases} \frac{2\kappa + \lambda(1 - 3n_1^2)}{4\pi\varepsilon^2}, \\ \frac{\kappa + \lambda(3n_3^2 - 1)}{\varepsilon\sqrt{2\pi}}, \\ \kappa - \lambda, \end{cases} \quad \eta = -3\lambda \begin{cases} \frac{3n_1^2 - 1}{8\varepsilon\sqrt{2\pi}}, \\ 1/2, \\ 1, \end{cases} \quad L_{\mathbf{n}} = \begin{cases} \partial_{xx}, & d = 1, \\ \partial_{\mathbf{n}_\perp\mathbf{n}_\perp} - n_3^2\nabla^2, & d = 2, \\ \partial_{\mathbf{nn}}, & d = 3, \end{cases}$$

with $\kappa = \frac{4\pi N a_s}{x_s}$, $\lambda = \frac{mN\mu_0\mu_{\rm dip}^2}{3\hbar^2 x_s}$, $\varepsilon = \frac{1}{\sqrt{\gamma_z}}$, $\mathbf{n}_\perp = (n_1, n_2)^T$, and

$$G(\mathbf{x}) = \begin{cases} \frac{1}{\varepsilon\sqrt{2\pi}}\int_0^\infty \frac{e^{-s/2\varepsilon^2}}{\sqrt{s^2 + |\mathbf{x}|^2}}ds \\ 1/(2\pi|\mathbf{x}|), \\ \frac{1}{(2\pi)^{3/2}}\int_{\mathbb{R}} \frac{e^{-s^2/2}}{\sqrt{|\mathbf{x}|^2 + \varepsilon^2 s^2}}ds, \\ 1/(4\pi|\mathbf{x}|), \end{cases} \Leftrightarrow \widehat{G}(\xi) = \begin{cases} \frac{\varepsilon\sqrt{2}}{\sqrt{\pi}}\int_0^\infty \frac{e^{-\varepsilon^2 s/2}}{s + |\xi|^2}ds, & d = 1\&\text{SAM}, \\ 1/|\xi|, & d = 2\&\text{SDM}, \\ \frac{1}{2\pi^2}\int_{\mathbb{R}} \frac{e^{-\varepsilon^2 s^2/2}}{|\xi|^2 + s^2}ds, & d = 2\&\text{SAM}, \\ 1/|\xi|^2, & d = 3, \end{cases}$$

where $\widehat{f}(\xi)$ denotes the Fourier transform of a function $f(\mathbf{x})$ for $\mathbf{x}$, $\xi \in \mathbb{R}^d$. In addition, in 3D, u in (5.9) satisfies the Poisson equation [6, 7, 30]

$$-\nabla^2 u(\mathbf{x},t) = |\psi(\mathbf{x},t)|^2, \quad \mathbf{x} \in \mathbb{R}^3, \quad \text{satisfying} \lim_{|\mathbf{x}|\to\infty} u(\mathbf{x},t) = 0, \quad t \geq 0; \quad (5.10)$$

and in 2D with SDM approximation, u in (5.9) satisfies the fractional Poisson equation [6, 7, 30]

$$(-\nabla^2)^{1/2} u(\mathbf{x},t) = |\psi(\mathbf{x},t)|^2, \ \mathbf{x} \in \mathbb{R}^2, \ \text{satisfying} \lim_{|\mathbf{x}|\to\infty} u(\mathbf{x},t) = 0, \ t \geq 0. \quad (5.11)$$

The GPE (5.9) conserves the normalization (2.5) and energy per particle

$$E(\psi(\cdot,t)) = \int_{\mathbb{R}^d} \left[\frac{1}{2} |\nabla\psi|^2 + V(\mathbf{x})|\psi|^2 + \frac{\beta}{2} |\psi|^4 + \frac{\eta}{2}\varphi|\psi|^2 \right] d\mathbf{x} \equiv E(\psi(\cdot,0)), \ t \geq 0.$$

The ground state can be defined the same as (3.5) with the above energy functional. For the existence and uniqueness as well as nonexistence of the ground state of the dipolar BEC (5.9) and efficient and accurate numerical methods for simulation, such as BESP [10] or BEFP with nonuniform FFT [49], we refer to [6, 7, 10] and references therein. Similarly, for the well-posedness of the Cauchy problem of (5.9) with the initial data (4.1) and its dynamical properties as well as efficient and accurate numerical methods, such as TSSP [10] or TSFP with nonuniform FFT [49], we refer to [6, 7, 10] and references therein.

5.3. For spin-orbit-coupled BEC. At temperatures T much smaller than the critical temperature T_c, a spin-orbit-coupled BEC with two components can be well described by the macroscopic wave function $\Psi := \Psi(\mathbf{x},t) = (\psi_1(\mathbf{x},t), \psi_2(\mathbf{x},t))^T$ whose evolution is governed by the following 3D coupled Gross-Pitaevskii equations (CGPEs) [5, 7, 8, 48, 60, 65, 67, 73] for $\mathbf{x} \in \mathbb{R}^3$ and $t > 0$ as

$$i\hbar\partial_t\psi_1 = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) + \frac{i\hbar\tilde{k}_0}{2m}\partial_x + \frac{\hbar\tilde{\delta}}{2} + Ng_{11}|\psi_1|^2 + Ng_{12}|\psi_2|^2 \right] \psi_1 + \frac{\hbar\tilde{\Omega}}{2}\psi_2,$$
$$i\hbar\partial_t\psi_2 = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) - \frac{i\hbar\tilde{k}_0}{2m}\partial_x - \frac{\hbar\tilde{\delta}}{2} + Ng_{21}|\psi_1|^2 + Ng_{22}|\psi_2|^2 \right] \psi_2 + \frac{\hbar\tilde{\Omega}}{2}\psi_1,$$

where N is the total number of particles, $\tilde{k}_0$ describes the spin-orbit-coupling strength, $\tilde{\delta}$ is the detuning constant for Raman transition, $\tilde{\Omega}$ is the effective Rabi frequency describing the strength to realize the internal atomic Josephson junction (JJ) by a Raman transition, and the interactions of particles are described by $g_{jl} = \frac{4\pi\hbar^2 a_{jl}}{m}$ with $a_{jl} = a_{lj}$ $(j,l = 1,2)$ being the s-wave scattering lengths between the jth and lth components. The above CGPEs is normalized as

$$\|\Psi\|^2 := \int_{\mathbb{R}^3} \left[|\psi_1(\mathbf{x},t)|^2 + |\psi_2(\mathbf{x},t)|^2 \right] d\mathbf{x} = 1. \quad (5.12)$$

Again, under the harmonic potential (2.9), similarly to the nondimensionalization in section 2.2 and dimension reduction in 2.3 from 3D to 2D and 1D, we can obtain the following dimensionless CGPEs under the normalization condition (5.12) for spin-orbit-coupled BEC

in d-dimensions ($d = 1, 2, 3$) for $\mathbf{x} \in \mathbb{R}^d$ and $t > 0$ as

$$\begin{aligned} i\partial_t \psi_1 &= \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + ik_0\partial_x + \frac{\delta}{2} + \beta_{11}|\psi_1|^2 + \beta_{12}|\psi_2|^2 \right] \psi_1 + \frac{\Omega}{2}\psi_2, \\ i\partial_t \psi_2 &= \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) - ik_0\partial_x - \frac{\delta}{2} + \beta_{21}|\psi_1|^2 + \beta_{22}|\psi_2|^2 \right] \psi_2 + \frac{\Omega}{2}\psi_1, \end{aligned} \tag{5.13}$$

where $k_0 = \frac{\tilde{k}_0}{\omega_x}$, $\delta = \frac{\tilde{\delta}}{\omega_x}$, $\Omega = \frac{\tilde{\Omega}}{\omega_x}$, and β_{11}, $\beta_{12} = \beta_{21}$, β_{22} are dimensionless interaction constants. This CGPEs conserves the normalization (or total mass)

$$N(\Psi(\cdot,t)) := \|\Psi(\cdot,t)\|^2 = \int_{\mathbb{R}^d} \sum_{j=1}^{2} |\psi_j(\mathbf{x},t)|^2 \, d\mathbf{x} \equiv N(\Psi(\cdot,0)) = 1, \quad t \geq 0, \tag{5.14}$$

and the energy per particle

$$\begin{aligned} E(\Psi(\cdot,t)) = \int_{\mathbb{R}^d} \Bigg\{ \sum_{j=1}^{2} \left[\frac{1}{2}|\nabla\psi_j|^2 + |\psi_j|^2 \left(V(\mathbf{x}) + \frac{1}{2}\sum_{l=1}^{2} \beta_{jl}|\psi_l|^2 \right) \right] + \frac{\delta}{2}\left(|\psi_1|^2 - |\psi_2|^2 \right) \\ + ik_0\left(\overline{\psi}_1 \partial_x \psi_1 - \overline{\psi}_2 \partial_x \psi_2 \right) + \Omega\, \mathrm{Re}(\psi_1 \overline{\psi}_2) \Bigg\} d\mathbf{x} \equiv E(\Psi(\cdot,0)), \quad t \geq 0. \end{aligned} \tag{5.15}$$

In addition, when $\Omega = 0$, then it also conserves the mass of each component

$$N(\psi_j(\cdot,t)) := \int_{\mathbb{R}^d} |\psi_j(\mathbf{x},t)|^2 \, d\mathbf{x} \equiv N(\psi_j(\cdot,0)), \quad t \geq 0, \quad j = 1, 2. \tag{5.16}$$

The ground state can be defined as: Find $\Phi_g \in S$ such that

$$E_g := E(\Phi_g) = \min_{\Phi \in S} E(\Phi), \tag{5.17}$$

where $S = \{\Phi = (\phi_1, \phi_2)^T \mid \|\Phi\| = 1, \ E(\Phi) < \infty\}$. Of course, when $\Omega = 0$, for any fixed $0 \leq \alpha \leq 1$, an α-dependent ground state can be defined as: Find $\Phi_g^\alpha \in S_\alpha$ such that

$$E_g^\alpha := E(\Phi_g^\alpha) = \min_{\Phi \in S_\alpha} E(\Phi), \tag{5.18}$$

where $S_\alpha = \{\Phi = (\phi_1, \phi_2)^T \mid \|\phi_1\|^2 = \alpha, \|\phi_1\|^2 = 1 - \alpha, \ E(\Phi) < \infty\}$. It is easy to see that

$$E_g = E(\Phi_g) = \min_{0 \leq \alpha \leq 1} E_g^\alpha = \min_{0 \leq \alpha \leq 1} E(\Phi_g^\alpha) = \min_{0 \leq \alpha \leq 1} \min_{\Phi \in S_\alpha} E(\Phi). \tag{5.19}$$

For the existence and uniqueness as well as nonexistence of the ground states of the spin-orbit-coupled BEC (5.3) based on the definition (5.17) for any $\Omega \in \mathbb{R}$ and the definition (5.18) for $\Omega = 0$, and efficient and accurate numerical methods for simulation, such as BEFD or BESP [5, 7, 8], we refer to [5, 7, 8, 65, 67, 73] and references therein. Similarly, for the well-posedness of the Cauchy problem of (5.3) with the initial data $\Psi(\mathbf{x}, 0) = \Psi_0(\mathbf{x})$ and its dynamical properties as well as efficient and accurate numerical methods, such as TSSP [5, 7], we refer to [5, 7, 8, 65, 67, 73] and references therein. Finally, by setting

$\psi_1(\mathbf{x},t) = \varphi_1(\mathbf{x},t)e^{i(\omega t+k_0x)}$ and $\psi_2(\mathbf{x},t) = \varphi_2(\mathbf{x},t)e^{i(\omega t-k_0x)}$ with $\omega = \frac{\delta-k_0^2}{2}$ in the CGPEs (5.13), we obtain for $\mathbf{x} \in \mathbb{R}^d$ and $t > 0$

$$\begin{aligned} i\partial_t\varphi_1 &= \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + \delta + \beta_{11}|\varphi_1|^2 + \beta_{12}|\varphi_2|^2\right]\varphi_1 + \frac{\Omega}{2}e^{-i2k_0x}\varphi_2, \\ i\partial_t\varphi_2 &= \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + \beta_{21}|\varphi_1|^2 + \beta_{22}|\varphi_2|^2\right]\varphi_2 + \frac{\Omega}{2}e^{i2k_0x}\varphi_1. \end{aligned} \tag{5.20}$$

This CGPEs conserves the normalization (5.14) for any $\Omega \in \mathbb{R}$ and (5.16) when $\Omega = 0$ with ψ_j replaced by φ_j for $j = 1, 2$. It is very useful in designing the most efficient and accurate numerical methods for computing ground states and dynamics, such as BESP and TSSP [5, 7, 8]), especially for the box potential, comparing to the system (5.13).

6. Conclusions and future perspectives

Due to its massive relations and applications in many different areas, such as atomic, molecular and optical physics, quantum optics, condense matter physics and low temperature physics, the research on theoretical, experimental and computational studies of BEC has been started almost century ago and has grown explosively (or exponentially) since 1995. Up to now, rich and extensive research results have been obtained in experimental and theoretical understanding of ground states and dynamics of BEC. The research in this area is still very active and highly demanded due to the latest experimental and/or technological advances in BEC, such as spinor BEC [18, 22, 47, 51], BEC with damping terms [15] or impurities [50] or random potentials [63], degenerate Fermi gas [45], Rydberg gas [53], spin-orbit-coupled BEC [60], BEC at finite temperature [72], etc. These achievements have brought great challenges to AMO community, condensed matter community, and computational and applied mathematics community for modeling, simulating and understanding various interesting phenomenons related to BEC. It becomes more and more interdisciplinary involving theoretical, computational and experimental physicists and computational and applied mathematicians as well as pure mathematicians.

Acknowledgements. The author is grateful to Beijing Computational Science Research Center for its hospitality during the writing of this paper.

References

[1] Abo-Shaeer, J. R., Raman, C., Vogels J. M., and Ketterle, W., *Observation of vortex lattices in Bose-Einstein condensates*, Science **292** (2001), 476–479.

[2] Aftalion, A., *Vortices in Bose-Einstein Condensates*, Progress in Nonlinear Differential Equations and their Applications, 67, Birkhäuser, Boston, 2006.

[3] Anderson, M. H., Ensher, J. R., Matthewa, M. R., Wieman C. E., and Cornell, E. A., *Observation of Bose-Einstein condensation in a dilute atomic vapor*, Science **269** (1995), 198–201.

[4] Antoine, X., Bao, W., and Besse, C., *Computational methods for the dynamics of the nonlinear Schrödinger/Gross-Pitaevskii equations*, Comput. Phys. Commun. **184** (2013), 2621–2633.

[5] Bao, W., *Ground states and dynamics of multicomponent Bose-Einstein condensates*, Multiscale Model. Simul. **2** (2004), 210–236.

[6] Bao, W., Ben Abdallah N., and Cai, Y., *Gross-Pitaevskii-Poisson equations for dipolar Bose-Einstein condensate with anisotropic confinement*, SIAM J. Math. Anal. **44** (2012), 1713–1741.

[7] Bao, W. and Cai, Y., *Mathematical theory and numerical methods for Bose-Einstein condensation*, Kinet. Relat. Mod. **6** (2013), 1-135.

[8] ———, *Ground states of two-component Bose-Einstein condensates with an internal atomic Josephson junction*, East Asia J. Appl. Math. **1** (2010), 49–81.

[9] ———, *Optimal error estimates of finite difference methods for the Gross-Pitaevskii equation with angular momentum rotation*, Math. Comp. **82** (2013), 99-129.

[10] Bao, W., Cai, Y., and Wang, H., *Efficient numerical methods for computing ground states and dynamics of dipolar Bose-Einstein condensates*, J. Comput. Phys. **229** (2010), 7874–7892.

[11] Bao, W., Chern I-L., and Lim, F. Y., *Efficient and spectrally accurate numerical methods for computing ground and first excited states in Bose-Einstein condensates*, J. Comput. Phys. **219** (2006), 836–854.

[12] Bao, W. and Du, Q., *Computing the ground state solution of Bose-Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comput. **25** (2004), 1674–1697.

[13] Bao, W., Du, Q., and Zhang, Y., *Dynamics of rotating Bose-Einstein condensates and its efficient and accurate numerical computation*, SIAM J. Appl. Math. **66** (2006), 758–786.

[14] Bao, W., Jaksch, D., and Markowich, P. A., *Numerical solution of the Gross-Pitaevskii equation for Bose-Einstein condensation*, J. Comput. Phys. **187** (2003), 318–342.

[15] ———, *Three dimensional simulation of jet formation in collapsing condensates*, J. Phys. B: At. Mol. Opt. Phys. **37** (2004), 329–343.

[16] Bao, W., Le Treust, L., and Méhats, F., *Dimension reduction for the anistropic Bose-Einstein condensates in the strong interaction regime*, arXiv:1403.2884 (2014).

[17] Bao, W., Li, H. L., and Shen, J., *A generalized Laguerre-Fourier-Hermite pseudospectral method for computing the dynamics of rotating Bose-Einstein condensates*, SIAM J. Sci. Comput. **31** (2009), 3685–3711.

[18] Bao, W. and Lim, F. Y., *Computing ground states of spin-1 Bose-Einstein condensates by the normalized gradient flow*, SIAM J. Sci. Comput. **30** (2008), 1925–1948.

[19] Bao, W., Marahrens, D., Tang, Q., Zhang, Y., *A simple and efficient numerical method for computing the dynamics of rotating Bose-Einstein condensates via a rotating Lagrangian coordinate*, SIAM J. Sci. Comput. **35** (2013), A2671-A2695.

[20] Bao, W. and Shen, J., *A fourth-order time-splitting Laguerre-Hermite pseudospectral method for Bose-Einstein condensates*, SIAM J. Sci. Comput. **26** (2005), 2020–2028.

[21] Bao, W. and Wang, H., *An efficient and spectrally accurate numerical method for computing dynamics of rotating Bose-Einstein condensates*, J. Comput. Phys. **217** (2006), 612–626.

[22] ———, *A mass and magnetization conservative and energy-diminishing numerical method for computing ground state of spin-1 Bose-Einstein condensates*, SIAM J. Numer. Anal. **45** (2007), 2177–2200.

[23] Bao, W., Wang, H., and Markowich, P. A., *Ground, symmetric and central vortex states in rotating Bose-Einstein condensates*, Commun. Math. Sci. **3** (2005), 57–88.

[24] Baranov, M. A., *Theoretical progress in many-body physics with ultracold dipolar gases*, Phys. Rep. **464** (2008), 71-111.

[25] Ben Abdallah, N., Méhats, F., Schmeiser, C., and Weishäupl, R. M., *The nonlinear Schrödinger equation with a strongly anisotropic harmonic potential*, SIAM J. Math. Anal. **37** (2005), 189–199.

[26] Bloch, I., Dalibard, J., and Zwerger, W., *Many-body physics with ultracold gases*, Rev. Mod. Phys. **80** (2008), 885-964.

[27] Bose, S. N., *Plancks gesetz und lichtquantenhypothese*, Zeitschrift fr Physik **3** (1924), 178–181.

[28] Bradley, C. C., Sackett, C. A., Tollett, J. J., and Hulet, R. G., *Evidence of Bose-Einstein condensation in an atomic gas with attractive interaction*, Phys. Rev. Lett. **75** (1995), 1687–1690.

[29] Bruderer, M., Bao, W., and Jaksch, D., *Self-trapping of impurities in Bose-Einstein condensates: Strong attractive and repulsive coupling*, EPL **82** (2008), 30004.

[30] Cai, Y., Rosenkranz, M., Lei, Z., and Bao, W., *Mean-field regime of trapped dipolar Bose-Einstein condensates in one and two dimensions*, Phys. Rev. A **82** (2010), 043623.

[31] Caradoc-Davis, B. M., Ballagh, R. J., and Burnett, K., *Coherent dynamics of vortex formation in trapped Bose-Einstein condensates*, Phys. Rev. Lett. **83** (1999), 895–898.

[32] Carles, R., *Semi-Classical Analysis for Nonlinear Schrödinger Equations*, World Scentific, 2008.

[33] Carles, R., Markowich, P. A., and Sparber, C., *On the Gross–Pitaevskii equation for trapped dipolar quantum gases*, Nonlinearity **21** (2008), 2569–2590.

[34] Cerimele, M. M., Pistella, F., and Succi, S., *Particle-inspired scheme for the Gross-Pitaevski equation: An application to Bose-Einstein condensation*, Comput. Phys. Comm. **129** (2000), 82–90.

[35] Chen T. and Pavlović N., *Derivation of the cubic NLS and Gross-Pitaevskii hierarchy from many body dynamics in $d = 3$ based on space time norms*, Ann. H. Poincare, **15** (2014), 543–588.

[36] Chen X., *On the rigorous derivation of the 3D cubic nonlinear Schrödinger equation with a quadratic trap*, Arch. Rational Mech. Anal. **210** (2013), 365–408.

[37] Cornell, E. A. and Wieman, C. E., *Nobel Lecture: Bose-Einstein condensation in a dilute gas, the first 70 years and some recent experiments*, Rev. Mod. Phys. **74** (2002), 875–893.

[38] Dalfovo, F., Giorgini, S., Pitaevskii, L. P., and Stringari, S., *Theory of Bose-Einstein condensation in trapped gases*, Rev. Mod. Phys. **71** (1999), 463–512.

[39] Davis, K. B., Mewes, M. O., Andrews, M. R., van Druten, N. J., Durfee, D. S., Kurn, D. M., and Ketterle, W., *Bose-Einstein condensation in a gas of sodium atoms*, Phys. Rev. Lett. **75** (1995), 3969–3973.

[40] Einstein, A., *Quantentheorie des einatomigen idealen gases*, Sitzungsberichte der Preussischen Akademie der Wissenschaften **22** (1924), 261–267.

[41] _____, *Quantentheorie des einatomigen idealen gases, zweite abhandlung*, Sitzungsberichte der Preussischen Akademie der Wissenschaften **1** (1925), 3–14.

[42] Erdős, L., Schlein, B., and Yau, H. T., *Derivation of the Gross-Pitaevskii equation for the dynamics of Bose-Einstein condensate*, Ann. Math. **172** (2010), 291–370.

[43] Fetter, A. L., *Rotating trapped Bose-Einstein condensates*, Rev. Mod. Phys. **81** (2009), 647–691.

[44] Gerard, P., Markowich, P. A., Mauser, N. J., and Poupaud, F., *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math. **50** (1997), 321–377.

[45] Giorgini, S., Pitaevskii, L. P., and Stringari, S., *Theory of ultracold atomic Fermi gases*, Rev. Mod. Phys. **80** (2008), 1215–1274.

[46] Gross, E. P., *Structure of a quantized vortex in boson systems*, Nuovo. Cimento. **20** (1961), 454–457.

[47] Ho, T. L., *Spinor Bose condensates in optical traps*, Phys. Rev. Lett. **81** (1998), 742–745.

[48] Jaksch, D., Gardiner, S. A., Schulze, K., Cirac J. I., and Zoller, P., *Uniting Bose-Einstein condensates in optical resonators*, Phys. Rev. Lett. **86** (2001), 4733–4736.

[49] Jiang, S., Greengard, L., and Bao, W., *Fast and accurate evaluation of nonlocal Coulomb and dipole-dipole interactions via the nonuniform FFT*, SIAM J. Sci. Comput. (2014), to appear arXiv:1311.4120.

[50] Johnson, T. J., Bruderer, M., Cai, Y., Clark, S. R., Bao, W., and Jaksch, D., *Breathing oscillations of a trapped impurity in a Bose gas*, EPL **98** (2012), article 26001.

[51] Kawaguchi Y., and Ueda, M., *Spinor Bose-Einstein condensates*, Phys. Rep. **520** (2012), 253-381.

[52] Ketterle, W., *Nobel lecture: When atoms behave as waves: Bose-Einstein condensation and the atom laser*, Rev. Mod. Phys. **74** (2002), 1131–1151.

[53] Kiffner, M., Li, W., and Jaksch, D., *Three-body bound states in dipole-dipole interacting Rydberg atoms*, Phys. Rev. Lett. **111** (2013), 233003.

[54] Klainerman S. and Machedon M., *On the uniqueness of solutions to the Gross-Pitaevskii hierarchy*, Commun. Math. Phys. **279** (2008), 169–185 .

[55] Lahaye, T., Menotti, C., Santos, L., Lewenstein, M., and Pfau, T., *The physics of dipolar bosonic quantum gases*, Rep. Prog. Phys. **72** (2009) 126401.

[56] Leggett, A. J., *Bose-Einstein condensation in the alkali gases: Some fundamental concepts*, Rev. Mod. Phys. **73** (2001), 307–356.

[57] Lieb, E. H. and Seiringer, R., Derivation of the *Gross-Pitaevskii equation for rotating Bose gases*, Comm. Math. Phys. **264** (2006), 505–537.

[58] Lieb, E. H., Seiringer, R., Solovej, J. P., and Yngvason, J., *The Mathematics of the Bose Gas and its Condensation*, Oberwolfach Seminars 34, Birkhäuser Verlag, Basel, 2005.

[59] Lieb, E. H., Seiringer, R., and Yngvason, J., *Bosons in a trap: A rigorous derivation of the Gross-Pitaevskii energy functional*, Phys. Rev. A **61** (2000), 043602.

[60] Lin, Y. J., Jiménez-Garcia, K., and Spielman, I. B., *Spin-orbit-coupled Bose-Einstein condensates*, Nature **471** (2011), 83–86.

[61] London, F., *The λ-phenomenon of liquid helium and the Bose-Einstein degeneracy*, Nature **141** (1938), 643–644.

[62] Matthews, M. R., Anderson, B. P., Haljan, P. C., Hall, D. S., Wieman, C. E., and Cornell, E. A., *Vortices in a Bose-Einstein condensate*, Phys. Rev. Lett. **83** (1999), 2498–2501.

[63] Min, B., Li, T., Rosenkranz, M., and Bao, W., Subdiffusive spreading of a Bose-Einstein condensate in random potentials, *Phys. Rev. A* **86** (2012), article 053612.

[64] Minguzzi, A., Succi, S., Toschi, F., Tosi, M. P., and Vignolo, P., *Numerical methods for atomic quantum gases with applications to Bose-Einstein condensates and to ultracold fermions*, Phys. Rep. **395** (2004), 223-355.

[65] Pethick, C. J. and Smith, H., *Bose-Einstein Condensation in Dilute Gases*. Cambridge University Press, 2002.

[66] Pitaevskii, L. P., *Vortex lines in an imperfect Bose gas*, Soviet Phys. JETP **13** (1961), 451–454.

[67] Pitaevskii, L. P. and Stringari, S., *Bose-Einstein Condensation*, Clarendon Press, Oxford, 2003.

[68] Rosenkranz, M., Jaksch, D., Lim, F. Y., and Bao, W., *Self-trapping of Bose-Einstein condensates expanding in shallow optical lattices*, Phys. Rev. A **77** (2008) article 063607.

[69] Seiringer, R., *Gross-Pitaevskii theory of the rotating Bose gas*, Comm. Math. Phys. **229** (2002), 491–509.

[70] Sulem, C. and Sulem, P. L., *The Nonlinear Schrödinger Equation, Self-focusing and Wave Collapse*, Springer-Verlag, New York, 1999.

[71] Xiong, B., Gong, J., Pu, H., Bao, W., and Li, B., *Symmetry breaking and self-trapping of a dipolar Bose-Einstein condensate in a double-well potential*, Phys. Rev. A **79** (2009), 013626.

[72] Zaremba, E., Nikuni, T., and Griffin, A., *Dynamics of trapped Bose gases at finite temperature*, J. Low Temp. Phys. **116** (1999), 277.

[73] Zhang, Y., Bao, W., and Li, H. L., *Dynamics of rotating two-component Bose-Einstein condensates and its efficient computation*, Phys. D **234** (2007), 49–69.

Department of Mathematics, National University of Singapore, Singapore, 119076, Singapore
E-mail: matbaowz@nus.edu.sg; URL: http://www.math.nus.edu.sg/~bao/

Discrete-to-continuum variational methods for Lattice systems

Andrea Braides

Abstract. I review some recent results regarding the description of the behaviour of energy-driven discrete systems, and more precisely lattice systems, through the construction of approximate continuous problems. On one hand methods of weak convergence, homogenization, integral representation and gradient flow dynamics already used for continuum problems have been adapted to the discrete setting, on the other hand the new discrete dimension has brought new phenomena, novel problems and interesting results. I will limit my description to systems with interfacial energies, but focus on methods that can be adapted to a multi-scale analysis.

Mathematics Subject Classification (2010). Primary 49J45; Secondary 35B27, 35Q70, 49D50, 49F22, 49J55.

Keywords. Discrete systems, Lattice systems, variational methods, homogenization, optimal design, variational motion, Γ-convergence, gradient-flow dynamics, thin films.

1. Introduction

The presence of discrete systems is ubiquitous in the applications of Mathematics to Science and Technology, ranging from problems parameterized by the pixels of a screen in Computer Vision, to nodes on a network for Flow Dynamics, to the location of atoms in simulations of Continuum Mechanics problems, to that of larger ensembles in coarse-grained theories in Statistical Physics, etc. In many cases, the variables are directly parameterized on a *lattice*, or a portion of it (as in Image Processing or in the design of conducting networks), while in other cases this may be a simplifying assumption on the geometry or on the admissible states of the system. A paradigmatic example of the latter situation is the analysis of Lennard-Jones systems close to a ground state, for which a *crystallization* result holds; i.e., that minimal states can be parameterized on a regular lattice. Even this expected property of ground states is a very subtle issue and has been proved only in dimension two and for a class of interatomic potentials (see [61]).

We will consider only *variational* lattice systems; i.e., we will assume that their behaviour is governed by an energy functional depending on the values of a parameter u defined on the elements (nodes) of the lattice. We will mainly focus on a particular type of interactions, when the parameter u can take *only two values*, which we take being ± 1 (*spin variable*), the energy can be written as the sum of the interactions between pair of values of the parameter (*pair interactions*), and we will be interested in problems where the limit behaviour can be approximated by a continuous *surface energy*.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

It must be noted that the methods we will use are to some extent independent of the simplification that we are making. In many problems, surface energies appear only at some energy scale, or in competition with bulk terms (*free-discontinuity problems*). This is the case for example of models in Computer Vision (e.g., Blake and Zisserman's [15], whose continuous counterpart is the Mumford and Shah variational model [56]) or Lennard-Jones systems from which one can deduce continuous variational models in Fracture Mechanics [35]. Nevertheless, blow-up and localization techniques often allow to decouple surface and bulk contribution, and assume that the parameters are locally constant on both sides of interfaces, so that the results we are going to illustrate can be thought as describing a part of a technically more complex energy depending on a continuous parameter.

The main points of the presentation will be the following:

- a variety of techniques constructed for continuum energies such as *homogenization*, Γ-*convergence*, *multi-scale analysis*, *variational motions*, *Geometric Measure Theory*, are naturally suited for this discrete setting, and provide a natural environment to define a continuum approximation;
- conversely, the discrete dimension provides a much more natural environment where to state and solve some problems which in the continuum case can be stated only with complex topological and geometrical constructions;
- in some cases the choice of the parameters for the continuum description are the main difficulty. This choice gives different effects and provides new problems with respect to the continuum setting.

2. A variational setting for spin systems

We will fix a periodic lattice $\mathcal{L}$ in $\mathbb{R}^d$. For reasons of simplicity we will mainly think of $\mathbb{Z}^d$ or of the triangular lattice $\mathcal{T}$ in $\mathbb{R}^2$, but we may consider as well non-Bravais lattices such as the hexagonal lattice in $\mathbb{R}^2$, fcc or hcp lattices in $\mathbb{R}^3$, etc.

The energy setting. We will consider functions $u : \mathcal{L} \to \{1, -1\}$, whose value at $i \in \mathcal{L}$ is denoted by u_i, and pair-interaction energies defined on such functions. It will be sometimes convenient to write the energies in such a way that they are zero on the two constant states ± 1. Upon additive and multiplicative constants the general form of these functionals is

$$E(u) = \frac{1}{8} \sum_{ij} c_{ij}(u_i - u_j)^2. \tag{2.1}$$

The normalization constant $1/8$ is due to the fact that the pair (i, j) is accounted for twice and that $u_i - u_j \in \{-2, 0, 2\}$. It is more customary, especially in Statistical Physics, to write such energies as

$$E(u) = -\sum_{ij} c_{ij} u_i u_j, \tag{2.2}$$

which is an equivalent form if only a finite number of indices are taken into account (finite domain). If $c_{ij} \geq 0$ (*ferromagnetic interactions*) form (2.1) allows to directly consider an infinite domain and avoids $+\infty - \infty$ indeterminations. If interactions are not positive, it will be otherwise necessary to rewrite the energy in a different fashion with a suitable renormalization.

Convergence of discrete functions. We may use several notions of convergence of discrete functions. In most cases, given a family (u^ε) with $u^\varepsilon : \mathcal{L} \to \{\pm 1\}$ we consider functions $\widetilde{u}^\varepsilon$ which are piecewise-constant interpolations of the function with value u_i^ε on the node $\varepsilon i \in \varepsilon\mathcal{L}$. This correspond to scaling the lattice of a factor ε. The resulting functions belong to $L^1_{\rm loc}(\mathbb{R}^d)$. We can therefore consider the convergence $\widetilde{u}^\varepsilon \to u$ in this topology. In this way we have defined a *convergence $u^\varepsilon \to u$ of discrete functions to a continuum limit*. Other notions of convergence of $\widetilde{u}^\varepsilon$ to u will be used, and the corresponding convergences of u^ε to u.

Surface scaling and compactness. We will consider families of energies

$$E_\varepsilon(u) = \frac{1}{8}\sum_{ij} \varepsilon^{d-1} c_{ij}^\varepsilon (u_i - u_j)^2, \tag{2.3}$$

where in principle the sum runs over all pairs in $\mathbb{Z}^d \times \mathbb{Z}^d$ (with $i \neq j$). The scaling ε^{d-1} corresponds to considering $E_\varepsilon(u^\varepsilon)$ as surface energies if interpreted in the scaled parameter $\widetilde{u}^\varepsilon$. Indeed, in the simplest situation of *nearest-neighbour interactions* in the cubic lattice $\mathcal{L} = \mathbb{Z}^d$, with $c_{ij}^\varepsilon = 1$ if $|i-j| = 1$ and $c_{ij}^\varepsilon = 0$ otherwise, we have

$$E_\varepsilon(u^\varepsilon) = \mathcal{H}^{d-1}(\partial\{\widetilde{u}^\varepsilon = 1\}); \tag{2.4}$$

i.e., the energy coincides with the $d-1$-dimensional Hausdorff measure of the interface $\partial\{\widetilde{u}^\varepsilon = 1\}$, or, equivalently the perimeter of the set $\{\widetilde{u}^\varepsilon = 1\}$. From the theory of sets of finite perimeter, we deduce then that families with equibounded $E_\varepsilon(u^\varepsilon)$ are precompact with respect to the convergence $u^\varepsilon \to u$ (see e.g. [19]).

Static picture: Γ-limit. Functionals E_ε can be set in the framework of surface energies on sets of finite perimeter [11], and their behaviour is described by Γ-limits of the form

$$F(u) = \int_{\partial\{u=1\}} \varphi(x,\nu)\, d\mathcal{H}^{d-1}, \tag{2.5}$$

where ∂A denotes the *reduced boundary* of the set A, and ν its *measure-theoretical normal*.

In the homogeneous case $\varphi(x,\nu) = \varphi(\nu)$ this Γ-convergence will guarantee in particular the convergence (up to translations) of minimizers $\overline{u}^\varepsilon$ with the (limit) volume constraint $\#(\{\overline{u}^\varepsilon = 1\}) = M_\varepsilon$ and $M_\varepsilon\, \varepsilon^d \to 1$ to the function $\overline{u} = 2\chi_{\overline{A}} - 1$, where $\overline{A}$ is the *Wulff shape* of φ; i.e., $\overline{A}$ minimizes

$$A \mapsto \int_{\partial A} \varphi(\nu)\, d\mathcal{H}^{d-1}$$

among the sets of finite perimeter with unit volume, $|A| = 1$. Indeed the knowledge of the Wulff shape itself is sufficient to describe φ and hence the Γ-limit. In the simplest case (2.4) we have

$$\varphi(\nu) = \|\nu\|_1 := |\nu_1| + \cdots |\nu_d| \tag{2.6}$$

and $\overline{A}$ is a unit coordinate cube.

Dynamic picture: minimizing movement along a sequence of functionals. The knowledge of the Γ-limit does not give information sufficient to describe gradient-flow type dynamics, which may depend on the interaction between the *space scale* ε and the relevant

time scale τ, and is defined by an implicit-time discretization scheme along E_ε (see [23] and Section 5 below).

Note that for sufficiently slow time scales the behaviour of the systems is approximated by the gradient-flow dynamics related to the Γ-limit, which can therefore be used as a comparison motion. In case (2.6) and $d = 2$ the related dynamics is the *crystalline motion* of the sets $A_t := \{x : u(t,x) = 1\}$. In the case of an initial datum a square, the sets are all squares, with side length L satisfying the ODE

$$L' = -\frac{2}{L} \tag{2.7}$$

until extinction time [9].

3. Positive interactions

As remarked above, the simplest case for energies (2.1) is when all interactions are non-negative; in which case the only ground states are the uniform states. In that framework it is not restrictive, for the sake of notational simplicity, to limit the analysis to the cubic lattice $\mathcal{L} = \mathbb{Z}^d$. In the case of nearest-neighbour interactions; i.e., if $c_{ij} = 0$ if $|i - j| \neq 1$ energies E_ε can be directly rewritten as surface integrals. The discrete setting allows also to consider *long-range interactions*; i.e., interactions between non-neighbouring points.

3.1. Integral-representation results. A general question is whether an approximation by a surface energy can be used. The answer is positive if the decay of the interaction is sufficiently fast, as in the following theorem (where the hypotheses are simplified for the sake of simplicity of presentation).

Theorem 3.1 (compactness). *Let c^ε_{ij} be non-negative numbers satisfying*

(i) (coerciveness) $c^\varepsilon_{ij} \geq c_1 > 0$ *if* $|i - j| = 1$*;*

(ii) (decay) $|c^\varepsilon_{ij}| \leq c_2 |i - j|^{-r}$ *with* $r > d + 1$*,*

and let E_ε be defined by (2.3)*. Then, up to subsequences, E_ε Γ-converge to a surface energy of the form* (2.5) *for some Carathéodory integrand whose positively homogeneous extension of degree one in the second variable is convex. The domain of F is $BV_{\rm loc}(\mathbb{R}^d; \{\pm 1\})$.*

Remark 3.2.

(a) Conditions (i) and (ii) are simplified for expository reasons and can be easily improved;

(b) **(non-local limits)** if (ii) is relaxed to a growth condition only guaranteeing that F be finite on $BV_{\rm loc}(\mathbb{R}^d; \{\pm 1\})$, we may lose *locality*; e.g., F may be of the form

$$F(u) = \int_{\partial\{u=1\}} \varphi(x,\nu)\, d\mathcal{H}^{d-1} + \int_{\mathbb{R}^d \times \mathbb{R}^d} k(x,y)(u(x) - u(y))^2 \, dx\, dy; \tag{3.1}$$

(c) **(boundary terms)** in a finite domain Ω we can consider energies obtained by restricting the interactions in the definition of E_ε to i, j belonging to $\frac{1}{\varepsilon}\Omega$. If Ω is a sufficiently

smooth bounded set, then the corresponding limit takes the same form, but is restricted to functions in $BV(\Omega;\{\pm 1\})$ and takes into account only the part of $\partial\{u=1\}$ contained in Ω, up to a constant term concentrated on $\partial\Omega$. Note that this term may be relevant to problems where Ω is a *design parameter*.

Homogenization formulas. In the *periodic case*; i.e., if $c^\varepsilon_{ij} = C_{ij}$ and there exists $T \in \mathbb{N}$ such that $C_{ij} = C_{i+kT\ j+kT}$ for all $i,j,k \in \mathbb{Z}^d$ then $\varphi(x,\nu) = \varphi(\nu)$ and it is described by an asymptotic formula involving the computation of minimum problems with Dirichlet boundary conditions on a family of invading cubes [38]. The same result holds under *almost-periodicity assumptions*.

The use of homogenization formulas often allows to provide bounds on φ, and in some cases its actual computation. This is valid in particular when $T = 1$, so that $C_{ij} = C'_{i-j}$. For example, in the two-dimensional case with $C'_k = 1$ for $|k| \le \sqrt{2}$ and $C'_k = 0$ otherwise (*next-to-nearest neighbour interactions*) φ is described by its Wulff shape, which is an octagon.

We now give some examples in which the analysis of the effect of the discrete dimension in the homogenization formulas allows to highlight differences and new applications with respect to the analogous continuum problems.

3.2. Optimal design of networks. Lattice energies may be used to describe metric properties of networks. For the sake of simplicity we illustrate only a two-dimensional framework. In this case nearest-neighbour interaction energies on $\mathbb{Z}^2$ can be interpreted as a length functional on curves defined in the nodes of the translated lattice $(1/2,1/2)+\mathbb{Z}^2$, with piecewise-constant weights $a^\varepsilon_{(i+j)/2} = c^\varepsilon_{ij}$. The continuum counterpart of such energies are *Riemannian metrics* of the form

$$F_\varepsilon(\gamma) = \int_0^1 a^\varepsilon(\gamma(t))|\gamma'(t)|^2\,dt, \tag{3.2}$$

whose limits are given by *Finsler metrics* with the integrands $\varphi = \varphi(x,\nu)$ computed at $x = \gamma$ and $\nu = \gamma'$ [24, 29]. *Optimal-design problems* for such energies amount to finding the general form of such φ when a^ε are subject to some *design constraint*. The simplest such constraint is requiring that $a^\varepsilon \in \{\alpha,\beta\}$ where α and β are two positive constants (*mixture of two conductors*). In the continuum case such a description for φ is an open problem, with only bounds available [46]. The discrete case, where we require that $c^\varepsilon_{ij} \in \{\alpha,\beta\}$, allows a simple solution of this type of problems as follows.

A general observation (the "Dal Maso-Kohn localization principle" [22]) states that in order to describe general $\varphi(x,\cdot)$ it suffices to consider the case of periodic coefficients $C_{ij} \in \{\alpha,\beta\}$ with prescribed proportion of α and β connections (*microgeometries*). Often, the "extreme microgeometries" are then obtained by taking connections in parallel or in series. While this cannot be done in all directions at once in the continuum, such geometries are easily constructed in the discrete setting. As a result, reading the limit in terms of Finsler metrics, we obtain all functional of the form

$$F(\gamma) = \int_0^1 \varphi(\gamma(t),\gamma'(t))\,dt, \tag{3.3}$$

where $\varphi(x,\cdot)$ is any convex function satisfying

$$\alpha(|\nu_1|+|\nu_2|) \le \varphi(x,\nu) \le c_1|\nu_1|+c_2|\nu_2|,$$

where the coefficients c_1 and c_2 satisfy

$$c_1 \le \beta, \quad c_2 \le \beta, \qquad c_1+c_2 = 2(\theta\beta+(1-\theta)\alpha)$$

and θ is the limit local proportion of β connections at x.

3.3. Discrete thin objects. Theories of thin objects starting from three-dimensional bodies through a dimension-reduction procedure have been a very successful way to obtain rigorous simplified theories for membranes, shells, rods, etc. [31, 49, 54]. For elastic membranes, the three-dimensional energies have the form

$$F_\varepsilon(u) = \frac{1}{\varepsilon}\int_{D\times(0,\varepsilon)} W(\nabla u)\,dx_1\,dx_2\,dx_3\,, \tag{3.4}$$

and the resulting energies as $\varepsilon \to 0$ can be written on two-dimensional functions as

$$F(u) = \int_D \overline{W}(\nabla_\alpha u)\,dx_1\,dx_2, \tag{3.5}$$

where $\overline{W}$ is defined through a minimization and quasiconvexification procedure, and ∇_α denotes the gradient in dimension two. An analog description can be given for interfacial problems [30].

Similar energies can be considered in a discrete setting, simply restricting the summation in the definition of energies E_ε in (2.3) to a stripe

$$S_n^T = \{x \in \mathbb{R}^{d+1} : |\langle x,n\rangle| \le T\}, \tag{3.6}$$

where $T>0$ and $n \in S^d$. Note that the corresponding notion of convergence of function $u^\varepsilon \to u$ gives a limit function u defined on $\{x \in \mathbb{R}^{d+1} : |\langle x,n\rangle| = 0\}$, which we identify with $\mathbb{R}^d$. A compactness theorem, analogous to Theorem 3.1, guarantees, under the corresponding decay assumptions, that we have a limit functional of the form (2.5) defined in $\mathbb{R}^d$. Nevertheless, with respect to the continuum case, we have some notable differences.

1. Surface effects. Even in the simpler case of periodic C_{ij} and coordinate thin films; e.g., when $n = e_{d+1}$ is a coordinate vector, we have a non-trivial dependence of the resulting φ on T. This is due to the non-local character of discrete interactions, giving a boundary term which is predominant for small values of T.

2. Quasicrystals. When n is not rational (i.e., it is not a multiple of a vector in $\mathbb{Z}^{d+1}$) then the part of the lattice included in S_n^T cannot be considered as the superposition of copies of $\mathbb{Z}^d$. Nevertheless, almost-periodic techniques allow to cover also these cases, which are connected to the modeling of quasicrystals [28].

3. Aperiodic lattices. Penrose tilings. Some aperiodic lattices can be constructed through a "cut-and-project" procedure from a higher-dimensional lattice on a lower-dimensional subspace. This is the case of Penrose tilings on the plane, for example, which can be constructed

as a projection of a subset of $\mathbb{Z}^5$ on a suitable "irrational" two-dimensional subspace. To such a construction the techniques used for "quasicrystals" can be adapted, obtaining an effective interfacial energy [41].

Question. *Is the Wulff shape of such an interfacial energy a pentagon? How do these pentagons differ (depending on the corresponding Penrose tiling)?*

4. Objective structures. Nanotubes. The definition of homogenized interfacial energies can be extended from periodicity assumptions to *objective structures* [52]. As a particular case we can treat models of "brittle nanotubes", for which the effective interfacial energy can depend on the *chirality* of the model. It is interesting to note that, even though described by the same general formulas, the value of the fracture toughness depends very much on the type of underlying lattice considered.

3.4. Random models. Percolation. In a fashion connected to problems in Statistical Physics one can maintain a fixed lattice, and consider a random choice of coefficients. For simplicity we suppose that only nearest-neighbour interactions are taken into account and that $c_{ij}^{\varepsilon} \in \{\alpha, \beta\}$, in which case we can interpret this as a model of a uniform network with randomly placed defects or inclusions. In a two-dimensional framework (to which we limit our description), by the identification of boundaries with curves the limit can be interpreted as describing the overall metric properties of a random network. The precise statement requires the introduction of an i.i.d. random variable, which gives, for each of its realizations ω, a random choice of the coefficients $C_{ij}^{\omega} \in \{\alpha, \beta\}$ with

$$\begin{cases} C_{ij}^{\omega} = \beta & \text{with probability } p \\ C_{ij}^{\omega} = \alpha & \text{with probability } 1-p, \end{cases}$$

and in (2.3) we simply consider $c_{ij}^{\varepsilon} = C_{ij}^{\omega}$. In this way we obtain a family of functionals E_{ε}^{ω} indexed by the realizations of our random variable. The analysis of the Γ-limit for each fixed ω corresponds to fixing a distribution of connections through the realization ω of the random medium, and computing its overall properties, which in general may depend on ω; the question is if *almost surely* they do not (*deterministic limit*) and how can the limit be described in terms of known probabilistic quantities. In the case of $0 < \alpha \leq \beta < +\infty$ for each fixed ω the functionals are in the class taken into account by Theorem 3.1. Hence, a strictly positive and finite limit $\varphi = \varphi^{\omega}$ is always defined. Indeed, such a $\varphi^{\omega} = \varphi_p^{\alpha,\beta}$ is shown to almost surely depend only on the probability p and on the two values α and β, and can be described by the corresponding *first-passage percolation* formula [18, 38].

In the extreme cases, when $\alpha = 0$ or $\beta = +\infty$ (in this second case we use the convention that $+\infty \cdot 0 = 0$, so that $\beta(u_i - u_j)^2$ is finite, and equal to 0, if and only if $u_i = u_j$), the description of the limit is related to the properties of the *infinite clusters* of *bonds* (i.e., infinite connected components of elements of the dual lattice corresponding to pairs (i, j) with C_{ij}^{ω} taking the value α if $p < 1/2$, or, respectively, β if $p > 1/2$).

Theorem 3.3 (dilute-spin percolation theorem [37]). *Let $\alpha = 0$. Then we have*

(i) (trivial surface tension) *if $p \leq 1/2$ then the corresponding Γ-limit is almost surely 0 for all u;*

(ii) (non-trivial deterministic surface tension) *if $p > 1/2$ then there exists a homogeneous strictly positive surface tension $\varphi = \varphi_p^{0,\beta}$ such that* (2.5) *holds.*

The function $\varphi_p^{0,\beta}$ is given by the dilute first-passage percolation *formula* [44, 64].

For $p < 1/2$ the existence of an infinite cluster yields that the limit of minimal-length problems in the homogenization formula are always trivial. In the case $p > 1/2$ the discrete environment can be interpreted as a randomly perforated medium, for which the main issue is the estimate of the effect of 'large' holes (whose probability is very small), which is negligible thanks to the i.i.d. hypothesis.

Theorem 3.4 (rigid-spin percolation theorem [59]). *Let $\beta = +\infty$. Then we have*

(i) (non-trivial deterministic surface tension) *if $p \le 1/2$ then there exists a homogeneous strictly positive surface tension $\varphi = \varphi_p^{\alpha,\infty}$ such that* (2.5) *holds;*

(ii) (degenerate surface tension) *if $p > 1/2$ then the corresponding Γ-limit is almost surely equal to the functional identically $+\infty$ for all u, except for the trivial cases $u = 1$ and $u = -1$.*

The function $\varphi_p^{\alpha,\infty}$ is given by the chemical distance *on the strong cluster* [51]. *Furthermore, we have the continuity result $\varphi_p^{\alpha,\beta} \to \varphi_p^{\alpha,\infty}$ as $\beta \to +\infty$.*

The key point in this result is a variational characterization of the chemical distance (the asymptotic distance in the infinite cluster). It is interesting to note that the continuity result is not trivial, and relies on a lemma provided by H. Kesten [36], which we may state informally as follows.

Lemma 3.5. *Fixed $L < 1$, let γ be a long path in the infinite cluster with length less than L times the corresponding chemical distance between its endpoints. Then it must contain a fixed proportion $P = P(L)$ of connections in the complement of the infinite cluster.*

It is interesting and promising to note the mutual exchange of results and problems between variational results and the corresponding percolation techniques.

3.5. Some interesting generalizations. The hypotheses of the compactness Theorem 3.1 can be extended in several interesting directions.

1. Random lattices. The compactness Theorem 3.1 can be extended to random perturbations of a given lattice. Other cases than can be covered using similar techniques are Poisson processes. It would be very interesting to prove percolation results in this context.

2. Double-porosity models. Such models in the continuum case are used in applications, e.g., to the study of the flow in a naturally fractured reservoir [13, 26]. In a discrete setting the complex topological assumptions necessary to their modeling are simplified, and reduce to a degenerate coerciveness condition, where (i) in Theorem 3.1 is weakened to $c_{ij}^{\varepsilon} \ge \varepsilon$ for a part of the interactions. If we suppose that the part in which (i) remains valid in the strong form determines a finite family of N infinite periodic connected sets, each of which can be treated as a separate perforated set, then the limit depends on a N-dimensional variable $u \in BV_{\rm loc}(\mathbb{R}^d; \{\pm 1\})^N$, and takes the form

$$F(u) = \sum_{j=1}^{N} \int_{\partial\{u_j=1\}} \varphi_j(\nu_j)\, d\mathcal{H}^{d-1} + \int_{\mathbb{R}^d} \psi(u_1, \dots, u_N)\, dx,$$

where ψ is an interaction term. Note that we may add lower-order terms, which influence the form of ψ in a non-trivial way.

3. Energies depending on a finite number of parameters. Surfactants. The compactness theorem can be extended to functions taking finitely many values. Even in the simplest case of $u \in \{-1, 0, 1\}$ and supposing that we still have the trivial ground states ± 1, the effect of the extra variable can be described in detail by a different type of energy. As an example, in the *Blume-Emery-Griffith* model the 0-phase, suitably scaled, converges to a positive measure μ, and the limit can be written as

$$F(u, \mu) = \int_{\partial\{u=1\}} \phi\Big(\frac{d\mu}{d\mathcal{H}^{d-1}}, \nu\Big) d\mathcal{H}^{d-1} + c\|\mu\|(\mathbb{R}^d \setminus \partial\{u = 1\})$$

(see [7]). This functional describes the energetic effect of the 0-phase, which is different when we have "deposition" on the interface between the 1-phase and the -1-phase, or "dilution" on the interior of the phases. If we do not have constraints on μ, we may define $\varphi(\nu) = \min_{z\geq 0} \phi(z, \nu)$ and integrate out the measure variable, reobtaining a representation as in Theorem 3.1. Note that this optimization may not be possible if for example the total mass of μ is prescribed.

4. Interactions with negative or changing sign

When in (2.1) or (2.2) the interactions c_{ij}^{ε} may take a negative sign then formally the functional is not bounded below and then, in the case of the infinite domain, it must be suitably scaled, taking care to avoid $+\infty - \infty$ indeterminations [1]. Even after scaling in such a way that it becomes bounded from below, as in (2.3), the representation of the Γ-limit, and its computation, may be different than the one described in Theorem 3.1. As a simple example one can consider the triangular lattice and only the *anti-ferromagnetic* nearest-neighbour interactions; i.e., such that $c_{ij}^{\varepsilon} = -1$ when $|i - j| = 1$. In this geometry it is not possible that all pair interactions take the minimal value (*frustration*), so that minimal states are all those u not taking the same value at all vertices of each triangle. Note that in this case we may take the *magnetization* as the parameter for the Γ-limit, which is then trivially constant for all u with values in $[-1/3, 1/3]$ (the value $1/3$ corresponding to spins taking two values $+1$ and one -1 on each triangle, and analogously for $-1/3$), and no interfacial energy is present.

In the case of the cubic lattice $\mathbb{Z}^d$ the anti-ferromagnetic nearest-neighbour energies can be reduced, up to scaling, to ferromagnetic ones via a *change of parameter*, setting $v_i = (-1)^i u_i$ (where $(-1)^i = \Pi_k(-1)^{i_k}$). The phase boundaries for v are called *anti-phase boundaries* of u. Note that in this case the corresponding magnetization for u is always 0, both when $v = 1$ and $v = -1$. This shows that not always the magnetization is a good order parameter. The simple definition of v as above is not meaningful in general. For example, if we have nearest and next-to-nearest neighbour interactions in the square lattice with all antiferromagnetic interactions, such a change of variable gives an energy with still antiferromagnetic interactions, so that it cannot be rephrased as a ferromagnetic energy in terms of v. This shows that in general the determination of a limit order parameter is a non-trivial, and in general essential, part of the question.

4.1. Limits parameterized on ground states. The line followed in Theorem 3.1 can be repeated in this more general context if we can describe the *ground states* for the energy E. Actually, some care must be taken in the definition of ground states themselves. To that end, in the context of the cubic lattice, we assume to being able to rewrite our energies (upon

additive constants) as

$$E_\varepsilon(u) = \sum_i \varepsilon^{d-1} \Psi(\{u_{i+j}\}_{j\in\mathbb{Z}^d}),$$

and denote the energy localized to a set I by $E(u, I) = \sum_{i\in I} \Psi(\{u_{i+j}\}_{j\in\mathbb{Z}^d})$. Then the analog of Theorem 3.1 reads as follows.

Theorem 4.1. *Suppose that we have*

(i) (existence of periodic ground states) *there exist* $N, K \in \mathbb{N}$ *and* $\{v^1, \dots, v^K\}$ N*-periodic functions such that, setting* $Q_N = \{1, \dots, N\}^d$ *we have* $E(v^j, Q_N) = 0$*, and* $E(u, Q_N) \geq C > 0$ *if* $u \neq v^j$ *in* Q_N *for all* j*;*

(ii) (incompatibility of ground states) *if* $u = \begin{cases} v^l & \text{in } Q_N \\ v^m & \text{in } k + Q_N \end{cases}$ *with* $k \in \mathbb{Z}^d$ *such that* $Q_N \cap (k + Q_N) \neq \emptyset$*, then* $E(u, Q_N \cup (k + Q_N)) > 0$*;*

(iii) (decay of the energy) *if* $u = u'$ *in* $R\,Q_N$ *then* $|E(u', Q_N) - E(u, Q_N)| \leq C_R$ *and* $\sum_R C_R R^{d-1} < \infty$.

Then we have

(a) (compactness) *If* $E_\varepsilon(u_\varepsilon) \leq C < +\infty$*. Then there exist* $A_{1,\varepsilon}, \dots, A_{K,\varepsilon} \subseteq \mathbb{Z}^N$ *(identified with the union of the cubes centered on their points) such that* $u_\varepsilon = v^j$ *on* $A_{j,\varepsilon}$*, we have* $\chi_{A_{j,\varepsilon}} \to \chi_{A_j}$ *in the sense of convergence from discrete to continuum (Section* 2*), and* $A_1, \dots, A_K$ *is a partition of* $\mathbb{R}^d$*. We denote this convergence as*

$$u_\varepsilon \to u := \sum_{j=1}^K j\chi_{A_j};$$

(b) (Γ-convergence) *the* Γ*-limit can be written as*

$$F(u) = \int_{S(u)} \psi(u^+, u^-, \nu)\, d\mathcal{H}^{d-1}$$

where $S(u) := \bigcup_{i,j} \partial\{u = i\} \cap \partial\{u = j\}$ *and* $u^\pm$ *denote the traces of* u *on both sides of* $S(u)$*, for a suitable* BV-elliptic *function* ψ*.*

This result generalizes the previous compactness theorem provided that we enlarge our class of limit energies to *functionals defined on partitions of sets of finite perimeter* [11], and that we take the ground states themselves as *order parameters*. Conditions (ii) and (iii) have the same role as the hypotheses of Theorem 3.1.

In the case of nearest-neighbour antiferromagnetic energies on a cubic lattice, we have $K = N = 2$, with two 2-periodic ground states, corresponding to v^1 given by $v^1_i = (-1)^i$ and $v^2 = -v^1$.

4.2. Patterns. The parameterization of ground states can describe different *types of patterns* at the microscopic level. We list a few examples, where we do not explicit the translation that gives zero energy to ground states.

(1) for $d = 1$ and $E(u) = \sum_i (\alpha u_i u_{i-1} + u_{i-1} u_{i+1})$ with $|\alpha| < 2$ (strong anti-ferromagnetic next-to-nearest neighbour interactions in 1D) we have four 4-periodic ground states, differing by a translation, so that the order parameter can be interpreted as a *shift*;

(2) for $d = 2$ and $E(u) = c_1 \sum_{|i-j|=1} u_i u_j + c_2 \sum_{|k-l|=\sqrt{2}} u_k u_l$ in the square lattice, with $c_2 > 0$ and $2c_2 > c_1$ (strong anti-ferromagnetic next-to-nearest neighbour interactions in 2D) we have four 2-periodic ground states $v^1, \dots, v^4$ given by $v_i^1 = (-1)^{i_1}$, $v_i^2 = (-1)^{i_2}$, $v^3 = -v^1$, and $v^4 = -v^2$. In this case we have *striped* ground states. The two ground states v^1 and v^2 can be interpreted as the directions e_1 and e_2 while v^3 and v^4 as the opposite directions $-e_1$ and $-e_2$. The interface between v^1 and v^3, e.g., can be considered an *anti-phase boundary*. Note that the Wulff shape of $\psi(e_k, e_l, \cdot)$ can be either a square or an irregular hexagon [1];

(3) for $d = 2$ and $E(u) = c_1 \sum_{|i-j|=1} u_i u_j + c_2 \sum_{|k-l|=\sqrt{3}} u_k u_l$ in the triangular lattice, with $c_1 > 0$ and $c_2 < 0$ (anti-ferromagnetic nearest neighbour interactions and ferromagnetic next-to-nearest neighbour interactions) then we have six ground states, which are $\sqrt{3}$ periodic in the direction $(1/2, \sqrt{3}/2)$.

Remark 4.2 (boundary terms). Contrary to the ferromagnetic case, in general the additional boundary term appearing in problems on a bounded domain Ω is not a constant, and depends on the trace of the ground state at the boundary; i.e., the Γ-limit takes the form

$$F(u) = \int_{S(u)\cap\Omega} \psi(u^+, u^-, \nu)\, d\mathcal{H}^{d-1} + \int_{\partial\Omega} \phi(u, n) d\mathcal{H}^{d-1},$$

where n is the normal to $\partial\Omega$. As a consequence, minimizers can depend on the interplay between the geometry of the domain and the microstructure of ground states.

Remark 4.3 (ferromagnetic parameters). In some cases it is possible to infer, as in the case of dilute spin systems, that the relevant parameters can still be interpreted as the two ferromagnetic phases. This is the case of periodic antiferromagnetic inclusions, provided that the distance between the inclusion is sufficiently large with respect to their size. In this case the parameter indexing the ground states represents the *majority phase* in the ferromagnetic matrix. It is interesting to note that even in the two-dimensional setting the minimum problems giving the interfacial energy cannot be directly interpreted as minimal-length problems, as their computation may involve the value on large antiferromagnetic inclusions.

Remark 4.4 (change in parameters - open problems). Simple examples show that we may have a change in the limit parameter in dependence of the volume fraction between ferromagnetic and antiferromagnetic interactions. Optimal-design and random problems are widely open. For example, it is not proved whether *there exists a small but positive probability p such that a random i.i.d. distribution of antiferromagnetic interactions in a matrix of ferromagnetic interactions is still described by only two (ferromagnetic) states*.

4.3. Change of patterns in thin films. As we have observed in the ferromagnetic case, a boundary contribution can influence the limit description of discrete thin films. In that case, the influence was described by a varying value of the interfacial energy. In the case of the presence of antiferromagnetic interactions, taking into account that boundary conditions and the geometry influence the form of the ground states, we may have a more striking influence in dependence of the thickness of the thin film.

Remark 4.5 (variation of the parameters by boundary effects). We can take the two-dimensional antiferromagnetic/ferromagnetic Example (3) in Section 4.2, on a thin film with normal perpendicular to a direction of the lattice vectors. If we have a "very thin film" of a single layer, then we simply have a one-dimensional antiferromagnetic lattice, with 2 parameters. As we increase the number of layers, the number of parameters varies depending on the ratio c_1/c_2 and for a sufficiently large film it reaches the number of the parameter of the bulk case (i.e., 6).

Remark 4.6 (coerciveness by boundary effects). As we have remarked above, the antiferromagnetic triangular lattice is degenerate, with zero surface tension in the limit. In the case of thin films this is not the case. Not only, as above we have the one-dimensional antiferromagnetic lattice, with 2 parameters when we have a single layer, but we have a non degenerate interfacial energy for all number of layers, with the number of parameters that diverges as 2^N. Other interesting effects that can be highlighted on this simple example are the dependence on the thin film direction, and the asymmetry of the interfacial energy density.

5. Gradient-flow type dynamics

For energy-driven systems a notion of gradient-flow dynamics can be given through a time-discrete approximation scheme. For a sequence of parameterized energies F_ε defined on a Hilbert space X the *minimizing movement* $x(t)$ *along the sequence* F_ε with time scale $\tau = \tau(\varepsilon)$ from an initial datum x^0 (or from a family of initial data $x^0_\varepsilon \to x^0$) is defined as the limit of the (time-discrete) trajectories x^ε defined as follows: we set $x^\varepsilon_0 = x^0$ and define recursively x^ε_{k+1} as a solution of the minimum problem

$$\min\Big\{F_\varepsilon(x) + \frac{1}{2\tau}\|x - x^\varepsilon_k\|^2\Big\}, \tag{5.1}$$

and then $x^\varepsilon(t) = x^\varepsilon_{\lfloor x/\tau \rfloor}$ [12, 23]. This scheme can be adapted to discrete systems and energies of the type E_ε following a variant due to Almgren, Taylor and Wang [10] (which indeed precedes the formalization of minimizing movements). Note that the minimizing movement x depends on the time scale $\tau(\varepsilon)$. In particular we have the following extreme cases, under proper coerciveness assumptions, which are satisfied by our E_ε.

Theorem 5.1 (extreme minimizing movements). *Let F_ε be a sequence of functionals defined on discrete spaces. Let $x^0_\varepsilon \to x^0$ with $F_\varepsilon(x^0_\varepsilon) \le C < +\infty$ and let F_ε be equicoercive, non-negative and Γ-converge to F. Then*

(i) (pinning scale) *there exists a (sufficiently fast) scale τ_p such that if $\tau \le \tau_p$ then $x(t) = x^0$ for all t;*

(ii) (commutation scale) *there exists a (sufficiently slow) scale τ_c such that if $\tau \ge \tau_c$ then $x(t)$ coincides with the* minimizing movement *for F from x^0 (defined by taking $F_\varepsilon = F$).*

The existence of a pinning scale is a consequence of the discreteness of the space, and is independent of the Γ-convergence of F_ε. Loosely speaking, in the notation of (5.1) there exists a function f such that we have $\|x^0_\varepsilon - x\|^2 \ge f(\varepsilon)$ for all $x \ne x^0_\varepsilon$, so that the minimum in (5.1) is x^0_ε for all k if $\tau < f(\varepsilon)/2C =: \tau_p$.

The interesting regimes are those excluded by Theorem 5.1, which interpolate between the extreme scales. The relevant problems can be summarized as

(i) determine the *critical regime(s)* $\tau = \tau(\varepsilon)$ such that we neither have pinning nor commutation;

(ii) compute the corresponding continuum *effective minimizing movement*, and describe the additional features that make it differ from the "trivial" one of the commutative case.

This novel type of dynamic homogenization problems constitutes a very interesting and wide class of gradient-flow type dynamics. We only give a few examples in the case of discrete energies converging to a crystalline perimeter.

5.1. Homogenized dynamics for positive interactions. For many ferromagnetic (nearest-neighbour) interactions the Γ-limit F is given by the crystalline perimeter. In two dimensions, Almgren and Taylor have shown that the minimizing movement (*flat flow*) for this functional is given by motion by *crystalline curvature* [9]. This motion can be easily described for coordinate rectangles, in which case each side moves inwards with velocity given by its curvature κ, which in the crystalline case is defined by $\kappa = 2/L$, L being the length of the side. The same description holds for coordinate polyrectangles provided we define $\kappa = -2/L$ (i.e., the motion is outwards) if the set is concave at the side, and $\kappa = 0$ if the set is neither concave or convex at the side.

Remark 5.2 (partial pinning/quantization of the velocity [34]). If E_ε are ferromagnetic nearest-neighbour interactions with $c_{ij} = 1$ if $|i - j| = 1$, then we have

(i) the *critical regime* is $\varepsilon \sim \tau$;

(ii) if $\tau/\varepsilon \to \gamma$ then the *effective minimizing movement* is described by the law

$$v = \frac{1}{\gamma}\big\lfloor \gamma\kappa \big\rfloor,$$

with the convention on the crystalline curvature κ as above.

The integer part is explained by the fact that the "discrete sides" must move by a finite quantity (proportional) to ε. Note that, as a consequence, we have *partial pinning*; i.e., pinning of sides only when they are larger than 2γ, and that, contrary to the motion by crystalline curvature, we may have initial data which may be pinned after an initial motion.

Remark 5.3 (homogenization of the velocity [39]). As remarked in the case of the optimal design of ferromagnetic materials, we may have the same crystalline perimeter even when we have periodic inclusions with $c_{ij} = \beta > 1$ in a matrix of unit nearest-neighbour interactions. These inclusions do not influence the Γ-limit, but they do influence the resulting minimizing movement, the reason being that the "discrete sides" avoid the inclusions for energetic reasons. We still have an effective minimizing movement, with sides moving with a velocity

$$v = \frac{1}{\gamma} f_{\mathrm{hom}}\big(\gamma\kappa\big),$$

with γ as in the previous remark. The *homogenized velocity function* can be described through a homogenization formula, and takes into account the geometry and distribution of the strong inclusions.

Remark 5.4 (bulk effects). We can consider periodic inclusions as in the previous example but with $c^\varepsilon_{ij} = \varepsilon$ (*double-porosity scaling*). The effect of these inclusions is negligible in the Γ-limit, which can be treated as a perforated domain giving, upon properly scaling

the energy, still the same crystalline perimeter F. Nevertheless, in this case the effective minimizing movement has an additional term taking into account that the weak inclusions may be regarded as a bulk effect. We may have in the limit a velocity of the form

$$v = \frac{1}{\gamma}\big\lfloor \gamma c\kappa + c(\gamma)\big\rfloor.$$

Note that in this case the limit for very slow time scales is not the motion of the crystalline perimeter. The failure of Theorem 5.1(ii) is due to the non equi-coerciveness of E_ε in the double-porosity case. A similar type of effect is encountered in the study of convection in mushy layers [63].

5.2. Homogenized dynamics for non-positive interactions. We only highlight some phenomena in the case of antiferromagnetic interactions

Remark 5.5 (mobility and motion by creation of defects). In the case of multiple ground states the limit behaviour is connected to the motion of networks rather than sets. Even in the simplifying case when only two phases are present in the continuum description we may have

(1) (*mobility*) the velocity law may depend on the orientation of the boundary normal and on the two phases;

(2) (*motion by the creation of defects*) the interface may move by using an intermediate phase which is non optimal for static problems. This may also happen at corner points.

Remark 5.6 (motion by maximization). The discrete setting allows to define another kind of motion, e.g., by taking in the minimizing-movement scheme an antiferromagnetic interaction. In this way we formally define a motion following a *maximization criterion* of the ferromagnetic energy or a backward motion for the perimeter functional. In particular we may take as initial datum a single point (*nucleation*) from which we have an expanding family of sets at constant velocity, whose shape depends (on the energy and) on the distance [40].

6. Conclusion: surface scaling as part of a multi-scale analysis

For general discrete systems, the surface-energy description analyzed above must be placed in a proper multi-scale framework, together with effects related to other types of scaling. Note that, even when only energetic contributions are taken into account in a static picture described by a Γ-limit process, the same type of functionals can be considered with different scaling depending on the energy level. For the same quadratic energies we may have, e.g.,

(a) (*bulk scaling*) $\sum_{ij} \varepsilon^d c_{ij}|u_i - u_j|^2$ giving integral energies $\int f(x, u(x))\,dx$;

(b) (*surface scaling*) $\sum_{ij} \varepsilon^{d-1} c_{ij}|u_i - u_j|^2$ giving surface energies as described in this presentation above;

(c) (*vortex scaling*) $\sum_{ij} \varepsilon^{d-2}|\log\varepsilon|^{-1} c_{ij}|u_i - u_j|^2$ giving *vortex energies* defined on point singularities;

(d) (*gradient scaling*) $\sum_{ij} \varepsilon^{d-2} c_{ij} |u_i - u_j|^2$ giving integral energies depending on gradients $\int f(x, \nabla u(x))\, dx$, etc.

Such effects, and others, may be present at the same time. For some of them, methods corresponding to those described for surface energies have been developed and used. Some issues that have a direct link with the results described above are

- *vortex and dislocation models* from *vector spin systems* (the so-called *XY model*; i.e., with $u_i \in \mathbb{R}^2$ and $|u_i| = 1$) [4]. Even though the interactions are formally the same as the ferromagnetic one, here the relevant scaling gives a behaviour equivalent to *Ginzburg-Landau energies* both for the static and the dynamic case, and has applications in the theory of dislocations [6, 8];
- *liquid-crystal type models*. Here a very interesting issue is the choice of the parameter, which can be for example the *magnetization* or the *Q-tensor*, giving different limit theories even at the first bulk scaling [28];
- *microscopic order/disorder*. For computational and modelling reasons it is very important to know whether macroscopic energies correspond to a regular arrangements of discrete values (*Cauchy-Born rule*) or not, and how this 'regularity' properties depend on parameters (e.g. thickness of thin films) [2, 33, 47, 50, 58];
- *interaction between surface and bulk contribution for free-discontinuity problems*. Both in Computer Vision and Fracture Mechanics, among other applications, we encounter competing bulk and surface energies, which can be derived from atomistic Lennard-Jones interactions [35] or optimized among lattice energies (see e.g. [45]);
- *quasicontinuum methods*. Computational problems involving free-discontinuity problems for which details of interfacial interactions are important require a coupling between continuum discretization procedures and atomistic fine-mesh analysis (see e.g. [16, 57, 60]);
- *derivation of nonlinear and linear theories in Continuum Mechanics*. The interpretation of discrete energies as describing the deformation of a lattice ground state lead to the derivation of elastic energies. We may have nonlinear elastic energies even from very simple lattice interactions [3, 17], while linear elastic energies can be rigorously derived [42] using powerful rigidity estimates [48];
- *optimal-design problems*. As highlighted for surface energies, optimal design problems can be treated also for conducting or elastic networks [32] in the spirit of optimal design of composites [55]. Little has been done in this direction, which seems natural for applications;
- *surface relaxation and crystal shapes*. Surface energies may arise as a result of asymmetric interactions at internal or external boundaries as a higher-order effect [27, 62]. When other bulk interactions are close to a ground state an important effect of the surface energy is the determination of optimal shapes of crystals [14].

Among the general directions of research that have been taken, some that may be singled out are the elaboration of new notions in the direction to allow to bridge the scales (in the variational setting e.g. 'equivalence by Γ-convergence' [43]), the extension of variational techniques to non-zero temperature [53], and the removal of the assumption of an underlying lattice [14, 61]. Many more new questions have been raised, with a wide range of applications, but still, even in the simplest variational setting, many remain open.

Acknowledgements. The author is grateful to the Mathematical Institute of the University of Oxford for its hospitality during the writing of this paper.

References

[1] Alicandro, R., Braides, A., and Cicalese, M., *Phase and anti-phase boundaries in binary discrete systems: a variational viewpoint*, Netw. Heterog. Media **1** (2006), 85–107.

[2] _____, *Continuum limits of discrete thin films with superlinear growth densities*, Calc. Var. Partial Differential Equations **33** (2008), 267–297.

[3] Alicandro, R. and Cicalese, M., *A general integral representation result for continuum limits of discrete energies with superlinear growth*, SIAM J. Math. Anal. **36** (2004), 1–37.

[4] _____, *Variational analysis of the asymptotics of the XY model*, Arch. Ration. Mech. Anal. **192** (2009), 501–536.

[5] Alicandro, R., Cicalese, M., and Gloria, A., *Integral representation results for energies defined on stochastic lattices and application to nonlinear elasticity*, Arch. Ration. Mech. Anal. **200** (2011), 881–943.

[6] Alicandro, R., Cicalese, M., and Ponsiglione, M., *Variational equivalence between Ginzburg-Landau, XY spin systems and screw dislocations energies*, Indiana Univ. Math. J. **60** (2011), 171–208.

[7] Alicandro, R., Cicalese, M., and Sigalotti, L., *Phase transitions in presence of surfactants: from discrete to continuum*, Interfaces Free Bound. **14** (2012), 65–103.

[8] Alicandro, R., De Luca, L., Garroni, A., and Ponsiglione, M., *Metastability and dynamics of discrete topological singularities in two dimensions: a* Γ*-convergence approach*, Arch. Ration. Mech. Anal., to appear.

[9] Almgren, F. and Taylor, J. E., *Flat flow is motion by crystalline curvature for curves with crystalline energies*, J. Differential Geom. **42** (1995), 1–22.

[10] Almgren, F., Taylor, J. E., and Wang, L., *Curvature-driven flows: a variational approach*, SIAM J. Control Optim. **31** (1993), 387–438.

[11] Ambrosio, L. and Braides, A., *Functionals defined on partitions in sets of finite perimeter. I. Integral representation and* Γ*-convergence*, J. Math. Pures Appl. (9) **69** (1990), 285–305.

[12] Ambrosio, L., Gigli, N., and Savaré, G., *Gradient flows in metric spaces and in the space of probability measures. Second edition*, Lectures in Mathematics ETH Zrich. Birkhüser Verlag, Basel, 2008.

[13] Arbogast, T., Douglas, J. Jr., and Hornung, U., *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal. **21** (1990), 823–836.

[14] Au Yeung, Y., Friesecke, G., and Schmidt, B., *Minimizing atomic configurations of short range pair potentials in two dimensions: crystallization in the Wulff shape*, Calc. Var. Partial Differential Equations **44** (2012), 81–100.

[15] Blake, A. and Zisserman, A., *Visual reconstruction*, MIT Press Series in Artificial Intelligence. MIT Press, Cambridge, MA, 1987.

[16] Blanc, X., Le Bris, C., and Legoll, F., *Analysis of a prototypical multiscale method coupling atomistic and continuum mechanics*, M2AN Math. Model. Numer. Anal. **39** (2005), 797–826.

[17] Blanc, X., Le Bris, C., and Lions, P.-L., *From molecular models to continuum mechanics*, Arch. Ration. Mech. Anal. **164** (2002), 341–381.

[18] Boivin, D., *First passage percolation: the stationary case*, Probab. Th. Rel. Fields **86** (1990), 491-499.

[19] Braides, A., *Approximation of free-discontinuity problems*, Lecture Notes in Mathematics, 1694. Springer-Verlag, Berlin, 1998.

[20] _____, *Non-local variational limits of discrete systems*, Commun. Contemp. Math. **2** (2000), 285–297.

[21] _____, *Γ-convergence for beginners*, Oxford Lecture Series in Mathematics and its Applications, **22**. Oxford University Press, Oxford, 2002.

[22] _____, *A handbook of Γ-convergence*, In Handbook of Differential Equations, Stationary Partial Differential Equations, Volume 3 (M. Chipot and P. Quittner, eds.), Elsevier, 2006, pp. 101–213.

[23] _____, *Local minimization, variational evolution and Γ-convergence*, Lecture Notes in Mathematics, 2094. Springer, Cham, 2014.

[24] Braides, A., Buttazzo, G., and Fragalà, I., *Riemannian approximation of Finsler metrics*, Asymptot. Anal. **31** (2002), 177–187.

[25] Braides, A., Causin, A., and Solci, M., *Interfacial energies on quasicrystals*, IMA J. Appl. Math. **77** (2012), 816–836.

[26] Braides, A., Chiadò Piat, V., and Piatnitski, A., *A variational approach to double-porosity problems*, Asymptot. Anal. **39** (2004), 281–308.

[27] Braides, A. and Cicalese, M., *Surface energies in nonconvex discrete systems*, Math. Models Methods Appl. Sci. **17** (2007), 985–1037.

[28] Braides, A., Cicalese, M., and Solombrino, F., *Q-tensor continuum energies as limits of head-to-tail symmetric spins systems*, preprint 2013, to appear.

[29] Braides, A. and Defranceschi, A., *Homogenization of multiple integrals*, Oxford Lecture Series in Mathematics and its Applications, 12. Oxford University Press, New York, 1998.

[30] Braides, A. and Fonseca, I., *Brittle thin films*, Appl. Math. Optim. **44** (2001), 299–323.

[31] Braides, A., Fonseca, I., and Francfort, G., *3D-2D asymptotic analysis for inhomogeneous thin films*, Indiana Univ. Math. J. **49** (2000), 1367–1404.

[32] Braides, A. and Francfort, G. A., *Bounds on the effective behaviour of a square conducting lattice*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **460** (2004), 1755–1769.

[33] Braides, A. and Gelli, M.S., *Limits of discrete systems with long-range interactions. Special issue on optimization* (Montpellier, 2000), J. Convex Anal. **9** (2002), no. 2, 363–399.

[34] Braides, A., Gelli, M.S., and Novaga, M., *Motion and pinning of discrete interfaces*, Arch. Ration. Mech. Anal. **195** (2010), 469–498.

[35] Braides, A., Lew, A.J., and Ortiz, M., *Effective cohesive behavior of layers of interatomic planes*, Arch. Ration. Mech. Anal. **180** (2006), 151–182.

[36] Braides, A. and Piatnitski, A., *Overall properties of a discrete membrane with randomly distributed defects*, Arch. Ration. Mech. Anal. **189** (2008), 301–323.

[37] _____, *Variational problems with percolation: dilute spin systems at zero temperature*, J. Stat. Phys. **149** (2012), 846–864.

[38] _____, *Homogenization of surface and length energies for spin systems*, J. Funct. Anal. **264** (2013), 1296–1328.

[39] Braides, A. and Scilla, G., *Motion of discrete interfaces in periodic media*, Interfaces Free Bound. **15** (2013), 451–476.

[40] _____, *Nucleation and backward motion of discrete interfaces*, C. R. Math. Acad. Sci. Paris **351** (2013), 803–806.

[41] Braides, A. and Solci, M., *Interfacial energies on Penrose lattices*, Math. Models Methods Appl. Sci. **21** (2011), 1193–1210.

[42] Braides, A., Solci, M., and Vitali, E., *A derivation of linear elastic energies from pair-interaction atomistic systems*, Netw. Heterog. Media **2** (2007), 551–567.

[43] Braides, A. and Truskinovsky, L., *Asymptotic expansions by* Γ*-convergence*, Contin. Mech. Thermodyn. **20** (2008), 21–62

[44] Cerf, R. and Théret, M., *Law of large numbers for the maximal flow through a domain of* $\mathbb{R}^d$ *in first passage percolation*, Trans. Amer. Math. Soc. **363** (2011), 3665–3702.

[45] Chambolle, A., *Image segmentation by variational methods: Mumford and Shah functional and the discrete approximations*, SIAM J. Appl. Math. **55** (1995), 827–863.

[46] Davini, A. and Ponsiglione, M., *Homogenization of two-phase metrics and applications*, J. Anal. Math. **103** (2007), 157–196.

[47] E, W. and Ming, P., *Cauchy-Born rule and the stability of crystalline solids: static problems*, Arch. Ration. Mech. Anal. **183** (2007), 241–297.

[48] Friesecke, G., James, R.D., and Müller, S., *A theorem on geometric rigidity and the derivation of nonlinear plate theory from three-dimensional elasticity*, Comm. Pure Appl. Math. **55** (2002), 1461–1506.

[49] _____, Müller, S., *A hierarchy of plate models derived from nonlinear elasticity by Gamma-convergence*, Arch. Ration. Mech. Anal. **180** (2006), 183–236.

[50] Friesecke, G. and Theil, F., *Validity and failure of the Cauchy-Born hypothesis in a two-dimensional mass-spring lattice*, J. Nonlinear Sci. **12** (2002), 445–478.

[51] Garet, O. and Marchand, R., *Large deviations for the chemical distance in supercritical Bernoulli percolation*, Annals of Probability **35** (2007), 833–866.

[52] James, R. D., *Objective structures*, J. Mech. Phys. Solids **54** (2006), 2354–2390.

[53] Kotecký, R. and Luckhaus, S., *Nonlinear elastic free energies and gradient Young-Gibbs measures*, Commun. Math. Phys. (2014), to appear.

[54] Le Dret, H. and Raoult, A., *The nonlinear membrane model as variational limit of nonlinear three-dimensional elasticity*, J. Math. Pures Appl. (9) **74** (1995), 549–578.

[55] Milton, G.W., *The theory of composites.* Cambridge Monographs on Applied and Computational Mathematics, 6. Cambridge University Press, Cambridge, 2002.

[56] Mumford, D. and Shah, J., *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math. **42** (1989), 577–685.

[57] Ortner, C. and Süli, E., *Analysis of a quasicontinuum method in one dimension*, M2AN Math. Model. Numer. Anal. **42** (2008), 57–91.

[58] Schmidt, B., *On the passage from atomic to continuum theory for thin films*, Arch. Ration. Mech. Anal. **190** (2008), 1–55.

[59] Scilla, G., *Variational problems with percolation: rigid spin systems*, Adv. Math. Sci. Appl. **23** (2013), 187–207

[60] Tadmor, E.B. and Ortiz, M., *Phillips, R., Quasicontinuum analysis of defects in solids*, Philosophical Magazine A **73** (1996), 1529–1563

[61] Theil, F. *A proof of crystallization in two dimensions*, Comm. Math. Phys. **262** (2006), 209–236.

[62] _____, *Surface energies in a two-dimensional mass-spring model for crystals*, ESAIM Math. Model. Numer. Anal. **45** (2011), 873–899.

[63] Worster, M.G., *Convection in mushy layers*, Annu. Rev. Fluid Mech. **29** (1997), 91–122.

[64] Wouts, M., *Surface tension in the dilute Ising model*, The Wulff construction, Comm. Math. Phys. **289** (2009) 157–204.

Dipartimento di Matematica, Università di Roma 'Tor Vergata', 00185 Roma, Italy
E-mail: braides@mat.uniroma2.it

Mathematical models and numerical methods for electronic structure calculation

Eric Cancès

Abstract. This contribution provides a pedagogical introduction for mathematicians to the field of electronic structure calculation. The N-body electronic Schrödinger equation and the main methods to approximate the solutions to this equation (wavefunction methods, density functional theory, quantum Monte Carlo) are presented. The numerical simulation of the resulting models, the construction of electronic structure models for systems with infinitely many electrons (perfect crystals, crystals with local defects, disordered materials) by means of thermodynamic limits, and the modeling and simulation of molecules interacting with complex environments, are discussed.

Mathematics Subject Classification (2010). Primary 81Q05; Secondary 35A15.

Keywords. Schrödinger equation, variational methods, numerical methods, quantum chemistry, solid state physics, materials science.

1. Introduction

Electronic structure calculation has become an essential tool in chemistry, condensed matter physics, molecular biology, materials science, and nanosciences. Over 10,000 research articles containing electronic structure calculations were published in 2013, and the field utilizes about 15% of the CPU time available in scientific computing centers worldwide. Its importance in contemporary research was acknowledged by the 1998 Nobel Prize in Chemistry shared by Kohn and Pople for their contributions to density functional theory and wavefunction methods for electronic structure calculation. The 2013 Nobel Prize in Chemistry was then awarded to Karplus, Levitt and Warshel for the development of multiscale models for complex chemical systems, the finer scale being dealt with electronic structure models.

The field is also an inexhaustible source of exciting mathematical and numerical problems. In this contribution, we explain how to model and simulate a sample of matter at the molecular scale from the first principles of quantum mechanics. The central object is the time-independent N-body Schrödinger equation allowing one to compute the possible electronic states of the system for a given configuration of the nuclei. This equation is a linear elliptic eigenvalue problem on $\mathbb{R}^{3N}$ where N is the number of electrons in the system. The construction of mathematical approximations of this equation and the design of efficient numerical methods to directly simulate the 3N-dimensional linear Schrödinger equation and its various lower-dimensional but nonlinear approximations (such as the Hartree-Fock and Kohn-Sham models) is a very active field of interdisciplinary research. After introducing the

Proceedings of the International Congress of Mathematicians, Seoul, 2014

electronic Schrödinger equation and the main three approaches to approximate and simulate it (wavefunction methods, density functional theory, and quantum Monte Carlo methods), we present recent advances in the mathematical understanding of electronic structure models and in the methods to solve them numerically, as well as open questions and future research directions. In order to make this contribution self-contained, we first recall the fundamental principles of quantum mechanics and briefly discuss the Born-Oppenheimer approximation allowing one (in most cases) to decouple electronic and nuclear degrees of freedom.

2. Basics of (non-relativistic) quantum mechanics

An autonomous quantum system is described by a separable complex Hilbert space $\mathcal{H}$ called the state space, and a self-adjoint operator H on the state space called the Hamiltonian. The (pure) states of the system are in one-to-one correspondence with the projective space $\mathcal{H}/\mathbb{C}$, which means that the state of the system at time t is completely characterized by a normalized vector $\psi(t) \in \mathcal{H}$, called the wave function, $\psi(t)$ and $e^{i\alpha}\psi(t)$ describing the same state for all $\alpha \in \mathbb{R}$. The dynamics of the system is governed by the time-dependent Schrödinger equation

$$i\hbar \frac{d\Psi}{dt}(t) = H\Psi(t), \tag{2.1}$$

where $\hbar$ is the reduced Planck constant. Of particular importance are the stationary states, namely the states of the form $\Psi(t) = e^{i\alpha(t)}\psi$, where $\|\psi\|_{\mathcal{H}} = 1$ and where $e^{i\alpha(t)}$ is an irrelevant global phase factor. Inserting this Ansatz in the time-dependent Schrödinger equation (2.1), we obtain that there exists a real number E such that ψ satisfies the time-independent Schrödinger equation

$$H\psi = E\psi$$

and $\alpha(t) = -iEt/\hbar$. In other words, ψ is an eigenvector of the self-adjoint operator H associated with the eigenvalue E. From a physical viewpoint, E is the energy of the stationary state ψ.

The formalism presented above is completely general and valid for any isolated quantum system. For a simple system consisting of a single particle of mass m and spin s subjected to a stationary external potential $V_{\rm ext}$ (take for instance $V_{\rm ext} \in L^2(\mathbb{R}^3) + L^\infty(\mathbb{R}^3)$ to avoid technical problems), the state space is the Hilbert space $L^2(\mathbb{R}^3 \times \Sigma, \mathbb{C}) \equiv L^2(\mathbb{R}^3, \mathbb{C}^{2s+1})$, where Σ is a finite set of cardinality $2s+1$. In the so-called position representation, the integrable function $x \mapsto |\Psi(t;x,\sigma)|^2$ is the probability density of observing at time t the particle at point x with spin σ. The Hamiltonian is the self-adjoint operator on $L^2(\mathbb{R}^3 \times \Sigma, \mathbb{C})$ defined by

$$H = -\frac{\hbar^2}{2m}\Delta + V_{\rm ext},$$

where Δ is the Laplace operator with respect to the variable x, and $V_{\rm ext}$ the operator of multiplication by the real-valued function $V_{\rm ext}$. The first term of the Hamiltonian models the kinetic energy of the particle, and the second term its potential energy. In the absence of external magnetic field, an assumption we make here, both terms are independent of the spin variable, so that it is not necessary for our purpose to specify the physical meaning of the spin variable. Lastly, the time-evolution of the particle is driven by the linear time-dependent

Schrödinger equation in the three dimensional space:

$$i\hbar\frac{\partial\Psi}{\partial t}(t,x,\sigma) = -\frac{\hbar^2}{2m}\Delta\Psi(t,x,\sigma) + V_{\text{ext}}(x)\Psi(t,x,\sigma).$$

Likewise, the stationary states are obtained by solving a linear elliptic eigenvalue problem in the three dimensional space:

$$-\frac{\hbar^2}{2m}\Delta\psi(x,\sigma) + V_{\text{ext}}(x)\psi(x,\sigma) = E\psi(x,\sigma).$$

The spectrum of the Hamiltonian H obviously depends on the external potential V_{ext}. For typical potentials encountered in electronic structure calculation, it has the structure displayed on Fig. 2.1. The lowest energy eigenmode is called the ground state; the higher energy eigenmodes are called excited states.

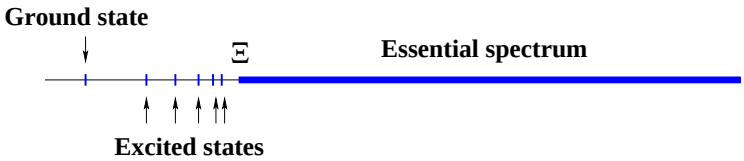


Figure 2.1. Typical spectra of the Hamiltonians encountered in (non-relativistic) electronic structure calculation of atoms and molecules: the essential spectrum is a half-line $[\Xi, +\infty)$; the discrete spectrum (bound states) can be empty, or consist of a finite or countable sequence of isolated eigenvalues of finite multiplicities. In the case when the number of discrete eigenvalues is infinite, they accumulate at Ξ. The essential spectrum coincides with the continuous spectrum (diffusion states) but can also contain eigenvalues embedded in the continuous spectrum.

Let us now consider a quantum system consisting of two particles of spins s_1 and s_2 respectively. The state space then is a subspace of the tensor product of the one-particle state spaces of the two particles. In other words, the wavefunction $\Psi(t)$ is a function of $L^2(\mathbb{R}^3 \times \Sigma_1, \mathbb{C}) \otimes L^2(\mathbb{R}^3 \times \Sigma_2, \mathbb{C}) \equiv L^2(\mathbb{R}^6, \mathbb{C}^{(2s_1+1)(2s_2+1)})$. In the position representation, the wavefunction Ψ is a function of the time variable t and of the position and spin variables $(x_i, \sigma_i) \in \mathbb{R}^3 \times \Sigma_i$ of each particle (where Σ_i is a set of cardinality $2s_i + 1$; for electrons, which are particles of spin $s = 1/2$, this set is usually denoted by $\{|\uparrow\rangle, |\downarrow\rangle\}$ and the spin states $|\uparrow\rangle$ and $|\downarrow\rangle$ are respectively called spin-up and spin-down). The function $|\Psi(t; x_1, \sigma_1; x_2, \sigma_2)|^2$ is interpreted as the probability density of observing at time t the particle 1 at point x_1 with spin σ_1 and the particle 2 at point x_2 with spin σ_2. If the two particles are different in nature (for instance an electron and a positron), the state space is equal to the tensor product $L^2(\mathbb{R}^3 \times \Sigma_1, \mathbb{C}) \otimes L^2(\mathbb{R}^3 \times \Sigma_2, \mathbb{C})$. It the two particles are identical bosons, the state space is the *symmetrized* tensor product $\mathcal{H} \vee \mathcal{H}$ of the one-particle state space $\mathcal{H} = L^2(\mathbb{R}^3 \times \Sigma, \mathbb{C})$, while if they are identical fermions, it is the *antisymmetrized* tensor product $\mathcal{H} \wedge \mathcal{H}$, so that the wavefunction Ψ must satisfy the symmetry properties

$$\Psi(t; x_2, \sigma_2; x_1, \sigma_1) = \Psi(t; x_1, \sigma_1; x_2, \sigma_2) \qquad \text{(for two identical bosons)},$$
$$\Psi(t; x_2, \sigma_2; x_1, \sigma_1) = -\Psi(t; x_1, \sigma_1; x_2, \sigma_2) \qquad \text{(for two identical fermions)}.$$

According to the spin-statistics theorem, particles with integer spins are bosons, while particles with half-integer spins are fermions. *In order to simplify the formalism, we will omit the spin variable in the sequel.*

Finally, for a quantum system consisting of N particles, the state space is a subspace of the tensor product of the one-particle state spaces and, in the position representation, $|\Psi(t;x_1,\cdots;x_N)|^2$ is the probability density of observing at time t the first particle at x_1, the second particle at x_2, etc. If all the N particles are identical, the state space is $\vee^N\mathcal{H}$ for bosons and $\wedge^N\mathcal{H}$ for fermions, and an important physical observable is the particle density

$$\rho_\Psi(t,x) = N\int_{\mathbb{R}^{3(N-1)}} |\psi(t;x,x_2,\cdots,x_N)|^2\,dx_2\cdots dx_N.$$

The time-independent Schrödinger equation is then a 3N-dimensional elliptic eigenvalue problem. In the case of N identical particles with mass m, subjected to a stationary external potential $V_{\rm ext}$, and interacting through a two-body potential W, we have

$$\left(-\sum_{i=1}^N \frac{\hbar^2}{2m}\Delta_{x_i} + \sum_{i=1}^N V_{\rm ext}(x_i) + \sum_{1\le i<j\le N} W(x_i,x_j)\right)\Psi(x_1,\cdots,x_N) = E\,\Psi(x_1,\cdots,x_N).$$

At first sight, it seems impossible to solve such an equation numerically for N greater than one or two. There is however one case, and this is key for electronic structure calculation, when this equation can be solved relatively easily. It is when the particles are identical and do not interact. In this case indeed, the Hamiltonian is separable (it is the sum of one-body operators)

$$H_{\rm NI} = -\sum_{i=1}^N \frac{\hbar^2}{2m}\Delta_{x_i} + \sum_{i=1}^N V_{\rm ext}(x_i) = \sum_{i=1}^N \mathfrak{h}_{x_i}$$

and the bound states of the N particle system can be easily obtained from the bound states $\phi_1,\phi_2,\cdots$ of the one-particle system, the latter being solutions to

$$\left\{\begin{array}{l} \mathfrak{h}\,\phi_i = \varepsilon_i\phi_i, \qquad \varepsilon_1\le\varepsilon_2\le\cdots \\ \displaystyle\int_{\mathbb{R}^3}\phi_i(x)\phi_j(x)\,dx = \delta_{ij}, \\ \mathfrak{h} = \displaystyle-\frac{\hbar^2}{2m}\Delta + V_{\rm ext}, \end{array}\right. \tag{2.2}$$

where $\varepsilon_1\le\varepsilon_2\le ...$ are the discrete eigenvalues of $\mathfrak{h}$ (counting multiplicities). In particular, the ground state of a system of N non-interacting bosons is given by

$$\Psi(x_1,\cdots,x_N) = \prod_{i=1}^N \phi_1(x_i), \quad \rho_\Psi(x) = N|\phi_1(x)|^2 \qquad \text{(non-interacting bosons)},$$

where ϕ_1 is the ground state eigenfunction of the one-body Hamiltonian $\mathfrak{h}$. For a system of N non-interacting fermions, the shape of the ground state wavefunction is more complex due to the antisymmetry constraint:

$$\Psi(x_1,\cdots,x_N) = \frac{1}{\sqrt{N!}}\det(\phi_i(x_j)), \quad \rho_\Psi(x) = \sum_{i=1}^N |\phi_i(x)|^2 \qquad \text{(non-int. fermions)},$$

where $(\phi_1,\cdots,\phi_N)$ is a family of L^2-orthonormal eigenfunctions of $\mathfrak{h}$ associated with the lowest N eigenvalues (counting multiplicities). Note that, as the Hamiltonian $H_{\rm NI}$ is real-valued, it is sufficient to consider real-valued wavefunctions.

3. First-principle molecular simulation

First principle molecular simulation is based on the simple observation that a given sample of matter is nothing but a collection of atomic nuclei and electrons in Coulomb interaction, and that we can specify the state space and the Hamiltonian of such a system provided we know its chemical formula (that is the chemical nature of the nuclei, and the number of electrons). Thus, the Hamiltonian of a molecular system composed of M nuclei with masses $m_1, \cdots, m_M$ and charges $z_1, \cdots, z_M$, and N electrons is

$$H_{\rm mol} = -\sum_{k=1}^{M} \frac{1}{2m_k}\Delta_{\mathbf{R}_k} - \sum_{i=1}^{N} \frac{1}{2}\Delta_{x_i} - \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{z_k}{|x_i - \mathbf{R}_k|} + \sum_{1\le i<j\le N} \frac{1}{|x_i - x_j|} + \sum_{1\le k<l\le M} \frac{z_k z_l}{|\mathbf{R}_k - \mathbf{R}_l|}.$$

The first term of $H_{\rm mol}$ models the kinetic energy of the nuclei, the second term, the kinetic energy of the electrons, and the last three terms the Coulomb interactions between nuclei and electrons, electrons and electrons, and nuclei and nuclei, respectively. Here and in the sequel, we adopt the system of atomic units obtained by setting to 1 the values of the reduced Planck constant $\hbar$, of the electron mass m_e, of the elementary charge e and of $4\pi\epsilon_0$, where ϵ_0 is the dielectric permittivity of the vacuum:

$$\text{atomic units:} \quad \hbar = 1, \quad m_e = 1, \quad e = 1, \quad 4\pi\epsilon_0 = 1.$$

A remarkable feature of this model is that it does not involve any empirical parameters specific to the system under consideration. It only depends on a few fundamental constants of physics, on the charges of the nuclei (1 for hydrogen, 2 for helium, 3 for lithium, ...), and on their masses, which have been measured experimentally with very high accuracy. This implies that the properties of the molecular system under consideration can be, *in principle*, computed from its chemical formula. The problem is that to do so, we need to solve the corresponding Schrödinger equation, that is a partial differential equation in dimension 3(N+M):

> *The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be solved.* (Dirac, 1929 [31])

Solving N-body Schrödinger equations for interacting particles was obviously not possible in Dirac's time, but it is doable nowadays, at least to some point. Before devoting a lot of effort to try and solve this equation for molecular systems, it is interesting to get some insights on the quality of the model. For this purpose, we first consider a very simple molecular system for which both extremely accurate calculations and experiments can be performed: a helium atom consisting of one nucleus of charge $z = 2$ and two electrons. If a laser with frequency ν is shone on a helium atom, and if ν is large enough, a photon with energy $h\nu$ is absorbed by the helium atom ($h = 2\pi\hbar$ is the Planck constant) causing an electron to escape to infinity (where it does not feel anymore the Coulomb potential generated by the nucleus and the other electron) with kinetic energy $E_{\rm k}$:

$$\mathrm{He} + \mathrm{h}\nu \;\rightarrow\; \mathrm{He}^{+} + \mathrm{e}^{-}.$$

The difference $\Delta E = h\nu - E_{\rm k} = h\Delta\nu$ between the energy of the photon and the kinetic energy of the electron at infinity, which can be measured very accurately, is equal to the difference between the ground state energy of the helium atom and the ground state energy of the He^+ ion (see Fig. 3.1), whose respective Hamiltonians are

$$H_{\rm He} = -\frac{1}{2m}\Delta_{\mathbf{R}} - \frac{1}{2}\Delta_{x_1} - \frac{1}{2}\Delta_{x_2} - \frac{2}{|x_1-\mathbf{R}|} - \frac{2}{|x_2-\mathbf{R}|} + \frac{1}{|x_1-x_2|},$$

and

$$H_{\rm He^+} = -\frac{1}{2m}\Delta_{\mathbf{R}} - \frac{1}{2}\Delta_{x_1} - \frac{2}{|x_1-\mathbf{R}|},$$

with $m \simeq 7294.29953$ a.u.

The ground state energy of He^+ can be computed analytically, and, using translational and rotational symmetries, the ground state energy of He can be obtained by solving numerically a 3-dimensional Schrödinger equation [55]. The agreement between theory ($\Delta\nu \simeq$ 5,945,262,288 MHz) and experiment (two independent experiments gave $\Delta\nu \simeq$ 5,945,204,238 MHz and $\Delta\nu \simeq$ 5,945,204,356 MHz) is very good, and even extremely good if the so-obtained solution is post-treated to include relativistic corrections by means of perturbation theory ($\Delta\nu \simeq$ 5,945,204,223 MHz). Note that relativistic effect cannot be dealt with perturbatively for heavy nuclei (see e.g. [35] for a mathematical analysis of fully relativistic quantum molecular models).

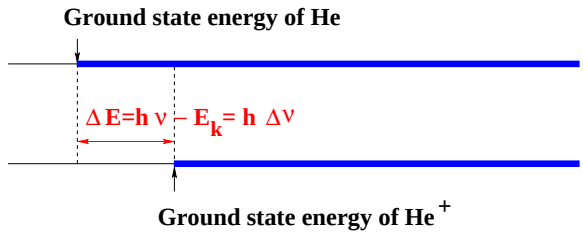


Figure 3.1. Spectra of He (top) and He^+ (bottom). The corresponding Hamiltonians being translation invariant, their discrete spectrum is empty.

3.1. Born-Oppenheimer approximation. Of course, the helium atom is a simple three body system. For more complex systems with dozens of nuclei and hundreds of electrons, approximations must be used. The first one is the so-called Born-Oppenheimer approximation. It relies on the fact that nuclei are much heavier than electrons. Using the mass ratio as a small parameter, it is possible to decouple the nuclear and electronic degrees of freedom by means of an adiabatic limit [77]. Using in a second stage a semiclassical approximation on the nuclear dynamics [1, 2], it is shown that, in most cases, nuclei behave as classical point-like particles interacting trough an effective potential energy function $W : \mathbb{R}^{3M} \to \mathbb{R}$ (see [12, 39] and references therein for mathematical studies of cases when the Born-Oppenheimer approximation breaks down). The state of the nuclei at time t is then described by the positions $\mathbf{R} = (\mathbf{R}_k(t))_{1\leq k\leq M} \in \mathbb{R}^{3M}$ and the momenta $\mathbf{P} = (\mathbf{P}_k(t))_{1\leq k\leq M}) \in \mathbb{R}^{3M}$ of the M nuclei, and the classical nuclear Hamiltonian reads

$$H_{\rm BO}(\mathbf{P},\mathbf{R}) = \sum_{k=1}^{M} \frac{|\mathbf{P}_k|^2}{2m_k} + W(\mathbf{R}_1,\cdots,\mathbf{R}_M).$$

The global minima of W are the equilibrium configurations of the molecular system. Loosely speaking, the local, non-global, minima correspond to metastable states and the critical points of W of Morse index 1 can be seen as the transition states of all the different possible chemical reactions or the conformational changes involving the atoms of the system. Sampling the nuclear configuration space according to suitable probability measures (note that the Gibbs measure $Z^{-1}e^{-\beta W}$ where β is an inverse temperature and Z a normalization constant is not well-defined for a molecule free to move in the whole space $\mathbb{R}^3$) allows one to compute the thermodynamical properties of the system [59]. The various dynamics which can be inferred from $H_{\rm BO}$ (classical Hamiltonian dynamics, Langevin dynamics, overdamped Langevin dynamics, ...) provide information on the kinetics of the reactions or conformational changes at thermodynamic equilibrium, as well as on non-equilibrium processes. The quantum counterpart of $H_{\rm BO}$ (obtained from $H_{\rm mol}$ by an adiabatic limit, not followed by a semiclassical limit) is also useful to compute properties such as the infrared spectra of molecules, proton tunneling in biological systems, or the superfluidity of helium 4 [28]. We will not further discuss the nuclear dynamics and focus on the electrons, which seem to have disappeared from the picture. They are in fact hidden in the definition of the effective potential W, the expression of which is given by

$$W(\mathbf{R}_1,\cdots,\mathbf{R}_M) = E_0^{\{\mathbf{R}_k\}} + \sum_{1\le k<l\le M} \frac{z_k z_l}{|\mathbf{R}_k-\mathbf{R}_l|}, \tag{3.1}$$

where $E_0^{\{\mathbf{R}_k\}}$ is the ground state eigenvalue of

$$H_{\rm elec}^{\{\mathbf{R}_k\}} = -\frac{1}{2}\sum_{i=1}^N \Delta_{x_i} - \sum_{i=1}^N\sum_{k=1}^M \frac{z_k}{|x_i-\mathbf{R}_k|} + \sum_{1\le i<j\le N}\frac{1}{|x_i-x_j|}.$$

The electronic Hamiltonian $H_{\rm elec}^{\{\mathbf{R}_k\}}$ is a self-adjoint operator on $\mathcal{H}_N := \wedge^N L^2(\mathbb{R}^3,\mathbb{C})$ (recall that we omit the spin variable) with domain $\mathcal{D}_N := \wedge^N H^2(\mathbb{R}^3,\mathbb{C})$ and form domain $\mathcal{Q}_N := \wedge^N H^1(\mathbb{R}^3,\mathbb{C})$.

3.2. Solving the electronic ground state problem. In the sequel, we focus on the computation of the ground state of the electronic Hamiltonian $H_{\rm elec}^{\{\mathbf{R}_k\}}$ for a given nuclear configuration $\{\mathbf{R}_k\}$. In order to simplify the notation, we will denote by $E_0 := E_0^{\{\mathbf{R}_k\}}$ and $H_N := H_{\rm elec}^{\{\mathbf{R}_k\}}$ (recall that N is the number of electrons in the system), so that

$$H_N = -\frac{1}{2}\sum_{i=1}^N \Delta_{x_i} + \sum_{i=1}^N V_{\rm ne}(x_i) + \sum_{1\le i<j\le N}\frac{1}{|x_i-x_j|} \quad\text{with}\quad V_{\rm ne}(x) = -\sum_{k=1}^M \frac{z_k}{|x-\mathbf{R}_k|}.$$

If the molecular system is neutral or positively charged, the spectrum of H_N has the structure sketched on Fig. 2.1, the number of discrete eigenvalues being infinite, and the bottom Ξ of the essential spectrum being the ground state of H_{N-1} [95]. In particular, $\Xi = 0$ for one electron systems, and $\Xi < 0$ if $N \ge 2$. The electronic ground state can be obtained by solving the minimization problem

$$E_0 = \inf\left\{\langle\Psi|H_N|\Psi\rangle,\ \Psi\in\mathcal{Q}_N,\ \|\Psi\|_{\mathcal{H}_N}=1\right\}, \tag{3.2}$$

where $\langle\psi|A|\psi\rangle$ denotes the quadratic form associated with the self-adjoint operator A (Dirac's bra-ket notation). Note that as $H_N\Psi$ is real-valued whenever Ψ is real-valued, it suffices to minimize over real-valued test functions Ψ.

Many approaches have been developed in the past 70 years to approximate electronic ground states, which can be classified into several families, the main ones being wavefunction methods [49], density functional theory [32] and quantum Monte Carlo methods (Fig. 3.2). All these methods have advantages and drawbacks. Some are more accurate than others, but require much higher computational effort, and their use is therefore limited to smaller systems. Some methods allow to correctly predict some properties (equilibrium geometry, infrared spectrum, polarizability, ...) but fail to predict other properties (reaction rates, magnetic shielding, ...). We refer e.g. to [49] for more details on these aspects.

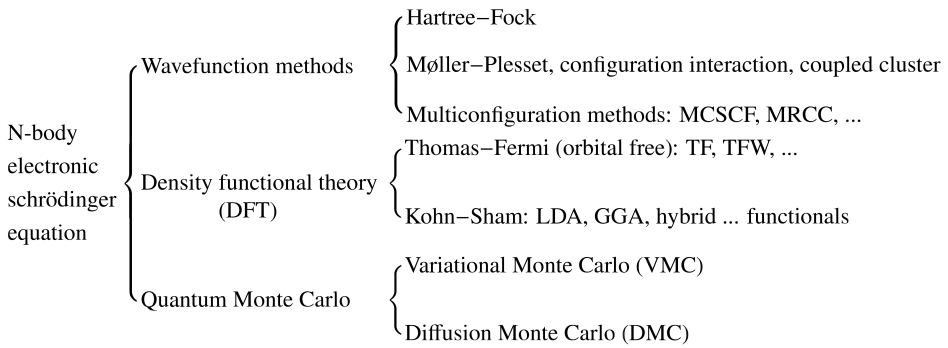


Figure 3.2. Classification of the main approximation methods for electronic ground state calculations.

The detailed description of all these methods goes beyond the scope of this pedagogical introduction. We will only explain the basic ideas underlying each of them. Beforehand, we report a numerical example illustrating the fact that electronic structure calculation can provide quantitatively correct results for polyatomic molecules. A water molecule H_2O consists of 3 nuclei and 10 electrons. The equilibrium geometry of the molecule (see Fig. 3.3) corresponds to the global minimizers of the interatomic potential W defined by (3.1). The results obtained with the Hartree-Fock model (93.96 pm, 106.33°) are reasonably close to experimental data. Better results are obtained with, for instance the Kohn-Sham LDA model [54] (96.86 pm, 105.00°), the Kohn-Sham GGA model with PBE functional [78] (96.90 pm, 104.75°), or the coupled cluster CCSD(T) method [49] (95.89 pm, 104.16°).

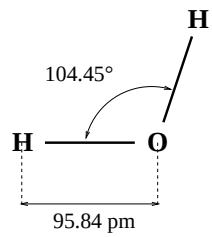


Figure 3.3. Geometry of the water molecule (experimental data).

3.2.1. Wavefunction methods. A natural way to approximate problem (3.2) is to minimize the energy functional $\langle\psi|H_N|\psi\rangle$ upon less general wavefunctions ψ. This leads to the

variational problem

$$E_0^{\mathcal{X}} = \inf \left\{ \langle \psi | H_N | \psi \rangle,\ \psi \in \mathcal{X},\ \|\psi\|_{\mathcal{H}} = 1 \right\}, \tag{3.3}$$

where $\mathcal{X}$ is a properly chosen subset of $\mathcal{Q}_N$. Obviously, $E_0^{\mathcal{X}}$ is an upper bound of the target value E_0. The Hartree-Fock (HF) approximation consists in minimizing over the set $\mathcal{X}$ of the L^2-normalized N-electron wavefunctions that can be written as an antisymmetrized product of single electron molecular orbitals ϕ_i:

$$\psi(x_1, \cdots, x_N) = \frac{1}{\sqrt{N!}} \det(\phi_i(x_j)). \tag{3.4}$$

Such functions are called *Slater determinants*. Recall that the ground state of a system of non-interacting fermions is a Slater determinant (see Section 2). Since the determinant is an alternate multilinear map, one can, without loss of generality, impose that the functions ϕ_i satisfy the orthonormality constraints

$$\int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij}.$$

When $\mathcal{X}$ is the set of Slater determinants, problem (3.3) can be rewritten, once the computation of $\langle \psi | H_N | \psi \rangle$ is explicitly performed, as

$$E_0^{\rm HF} = \inf \left\{ \mathcal{E}^{\rm HF}(\Phi),\ \Phi = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3))^N,\ \int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \right\}, \tag{3.5}$$

where

$$\begin{aligned} \mathcal{E}^{\rm HF}(\Phi) \quad = \quad & \frac{1}{2} \sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla \phi_i|^2 + \int_{\mathbb{R}^3} \rho_\Phi V_{\rm ne} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_\Phi(x) \rho_\Phi(y)}{|x-y|}\, dx\, dy \\ & - \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\gamma_\Phi(x,y)|^2}{|x-y|}\, dx\, dy, \end{aligned} \tag{3.6}$$

where the density matrix γ_Φ and the density ρ_Φ associated with the N-tuple $\Phi = (\phi_1, \cdots, \phi_N)$ are defined as

$$\gamma_\Phi(x,y) = \sum_{i=1}^{N} \phi_i(x) \phi_i(y), \qquad \rho_\Phi(x) = \gamma_\Phi(x,x) = \sum_{i=1}^{N} |\phi_i(x)|^2. \tag{3.7}$$

Note that the density ρ_Φ is in fact the density ρ_ψ of the N-body wavefunction ψ defined by (3.4). The last term in the HF energy functional (3.6) is called the exchange term. It has a purely quantum nature (it arises from the antisymmetry of the electronic wavefunction) and has no classical counterpart. The first-order optimality conditions associated with the constrained optimization problem (3.5) read, after some algebraic manipulation (see e.g. [68] for details), as

$$\left\{ \begin{array}{l} \Phi^0 = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3))^N, \\ \mathfrak{h}^{\rm HF}_{\Phi^0}\, \phi_i = \varepsilon_i \phi_i, \\ \displaystyle\int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij}, \\ \mathfrak{h}^{\rm HF}_{\Phi^0} \phi = -\dfrac{1}{2} \Delta \phi + V_{\rm ne} \phi + (\rho_{\Phi^0} \star |\cdot|^{-1}) \phi - \displaystyle\int_{\mathbb{R}^3} \frac{\gamma_{\Phi^0}(\cdot, x')}{|\cdot - x'|}\, \phi(x')\, dx'. \end{array} \right. \tag{3.8}$$

If Φ^0 is a minimizer of (3.5), then $\varepsilon_1, \cdots, \varepsilon_N$ are in fact the lowest N eigenvalues of the Hartree-Fock operator $\mathfrak{h}^{\rm HF}_{\Phi^0}$ (a stronger property is shown in [5]). These equations look somewhat similar to the equations (2.2) we obtained for non-interacting electrons. The big difference is that, this time, the mean-field Hamiltonian $\mathfrak{h}^{\rm HF}_{\Phi^0}$ depends on the ground state Φ^0. The Hartree-Fock equations (3.8) are therefore a *nonlinear* elliptic eigenvalue problem. The Hartree-Fock model has been thoroughly studied in the mathematical literature. The existence of a Hartree-Fock ground state for neutral molecules and positive ions was established by Lieb and Simon [68], and the set of solutions to the Hartree-Fock equations was studied by Lions [73]. Uniqueness is a difficult question [46]. Numerical algorithms are analyzed in [16, 22, 23, 60].

For polytatomic systems, the Hartree-Fock method is often not accurate enough to reach the requested accuracy. A natural way to improve the HF wavefunction $\Psi^{\rm HF}$ is to consider finite sums of Slater determinants and search for an approximation of the ground state wavefunction of the form

$$\Psi(x_1, \cdots, x_N) = c_0 \Psi^{\rm HF}(x_1, \cdots, x_N) + \text{finite sum of Slater determinants},$$

where c_0 is a normalization coefficient. Several ways to generate such improved wavefunctions have been proposed [49]: the Møller-Plesset perturbation method (mathematically based on Kato perturbation theory [51]), the configuration interaction method, and the coupled cluster method (see [82] for a mathematical analysis). The CCSD(T) method (the acronym stands for Coupled Cluster Single Double (Triple)) is considered as the gold standard of quantum chemistry. It however suffers from two major limitations:

- first, the required computational effort scales as N^7, instead of N^3 for the Hartree-Fock model, where N is the number of electrons in the system, so that its use is limited to relatively small molecular systems (a dozen of atoms); the construction of coupled cluster methods with better scaling is an active field of research;
- second, the coupled cluster method fails when the Hartree-Fock wavefunction is not the dominant component of the ground state wavefunction, that is when several Slater determinants are necessary to get a decent estimate of the ground state wavefunction.

To address the second problem, one has to resort to multi-configuration methods. These methods include the multi-configuration self-consistent field (MCSCF) method [49], which was mathematically analyzed by Le Bris [57], Friesecke [41] and Lewin [62], as well as the recently developed multi-reference coupled cluster (MRCC) method [83]. The MCSCF method can be interpreted as a low rank tensor method based on Tucker format [7, 37].

3.2.2. Density Functional Theory (DFT). It has been shown by Hohenberg and Kohn [50] that the electronic ground state energy and density can be obtained by minimizing a functional of the electronic density. To establish this fact, we report here the simple argument proposed later by Levy [61] (see also [64]). Splitting the electronic Hamiltonian as

$$H_N = H_N^1 + \sum_{i=1}^N V(x_i) \quad \text{with} \quad H_N^\lambda = -\frac{1}{2}\sum_{i=1}^N \Delta_{x_i} + \sum_{1\le i<j\le N} \frac{\lambda}{|x_i - x_j|},$$

and observing that, due to the antisymmetry property, we have for all $\Psi \in \mathcal{Q}_N$,

$$\langle \Psi | \sum_{i=1}^N V(x_i) | \Psi \rangle = \sum_{i=1}^N \int_{\mathbb{R}^{3N}} V(x_i) |\Psi(x_1, \cdots, x_N)|^2 \, dx_1 \cdots dx_N = \int_{\mathbb{R}^3} \rho_\Psi V,$$

the electronic ground state problem (3.2) also reads

$$E_0 = \inf\left\{\langle\Psi|H_N^1|\Psi\rangle + \int_{\mathbb{R}^3}\rho_\Psi V,\ \Psi\in\mathcal{Q}_N,\ \|\Psi\|_{\mathcal{H}_N}=1\right\}$$
$$= \inf\left\{F_N(\rho) + \int_{\mathbb{R}^3}\rho_\Psi V,\ \rho\in\mathcal{R}_N\right\},$$

where

$$\mathcal{R}_N = \{\rho \mid \exists\Psi\in\mathcal{Q}_N \text{ s.t. } \|\Psi\|_{\mathcal{H}_N}=1 \text{ and } \rho_\Psi=\rho\}$$

is the set of admissible electronic densities and where

$$F_N(\rho) = \inf\left\{\langle\Psi|H_N^1|\Psi\rangle,\ \Psi\in\mathcal{Q}_N,\ \|\Psi\|_{\mathcal{H}_N}=1,\ \rho_\Psi=\rho\right\} \tag{3.9}$$

is a *universal* density functional, in the sense that it only depends on the number of electrons in the system, not on the number, chemical natures, and positions of the nuclei. The set $\mathcal{R}_N$ can be easily characterized:

$$\mathcal{R}_N = \left\{\rho\geq 0 \mid \sqrt{\rho}\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\rho = N\right\}.$$

On the other hand, there is no simple way to evaluate $F_N(\rho)$ for a given $\rho\in\mathcal{R}_N$.

Although introduced decades before the works by Hohenberg and Kohn, Thomas-Fermi type models fall in the framework of DFT. They consist in approximating $F_N(\rho)$ by explicit functionals of the density ρ. Examples of such models include the original Thomas-Fermi (TF) model

$$F_{\rm TF}(\rho) = C_{\rm TF}\int_{\mathbb{R}^3}\rho^{5/3} + \frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy,$$

where $C_{\rm TF} = \frac{10}{3}(3\pi^2)^{2/3}$ is the Thomas-Fermi constant, and the Thomas-Fermi-von Weizsäcker (TFW) model

$$F_{\rm TFW}(\rho) = C_{\rm W}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2 + C_{\rm TF}\int_{\mathbb{R}^3}\rho^{5/3} + \frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy,$$

where the constant $C_{\rm W}$ takes different values depending on how the TFW model is derived [32]. The first term in $F_{\rm TF}(\rho)$ models the kinetic energy of the electrons ($C_{\rm TF}\rho_0^{5/3}$ is the kinetic energy density of a homogeneous gas of non-interacting electrons of uniform density ρ_0 [32]). The second component of $F_{\rm TF}(\rho)$ is the electrostatic energy of a *classical* charge distribution of density ρ. The first term in $F_{\rm TFW}(\rho)$ is a correction to the Thomas-Fermi approximation of the kinetic energy of the electrons taking into account the fact that, in molecular systems, the electronic density is not uniform. The above Thomas-Fermi type models provide crude approximations of $F_N(\rho)$ (according to the TF model, any molecule is unstable! [89]), and are no longer used in quantum chemistry and materials science. On the other hand, some improvements of the TFW model, the so-called *orbital-free models* [94], are used for the simulation of specific materials (aluminum crystals with defects for example [80]). The main reason why we mention Thomas-Fermi like models is that they are very useful in the mathematical analysis of electronic structure models [63, 67, 76]. They are indeed toy models upon which new mathematical techniques can be developed before being

applied to the more sophisticated models actually used in quantum chemistry and materials science.

The Kohn-Sham method [54] is to date the most popular electronic structure method as it provides the best compromise between computational efficiency and accuracy. It proceeds from the density functional theory remarking that on the one hand,

$$T_{\rm KS}(\rho) = \inf\left\{ \frac{1}{2}\sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla\phi_i|^2,\ \Phi = (\phi_i) \in (H^1(\mathbb{R}^3))^N,\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij},\ \rho_\Phi = \rho \right\},$$

where the density ρ_Φ is defined by (3.7), is an excellent approximation (exact for pure state non-interacting V-representable densities [64]) of the density functional for non-interacting electrons, that is of the density functional obtained by replacing H_N^1 with H_N^0 in (3.9), and that, on the other hand, the classical Coulomb energy

$$J(\rho) = \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy$$

is a reasonable first-order approximation of the electronic interaction. The functional $F_N(\rho)$ can therefore be decomposed as

$$F(\rho) = T_{\rm KS}(\rho) + J(\rho) + E_{\rm xc}(\rho), \tag{3.10}$$

where the exchange-correlation energy functional $E_{\rm xc}(\rho)$ is expected to be a small correction of the first two terms of the decomposition (3.10). Numerical simulations confirm that the exchange-correlation energy is about 10% of the total energy for ground states of molecular systems. The Kohn-Sham model then reads

$$E_0^{\rm KS} = \inf\left\{ \mathcal{E}^{KS}(\Phi),\ \Phi = (\phi_1,\cdots,\phi_N) \in (H^1(\mathbb{R}^3))^N,\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij} \right\}, \tag{3.11}$$

with

$$\mathcal{E}^{\rm KS}(\Phi) = \frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3} |\nabla\phi_i|^2 + \int_{\mathbb{R}^3} \rho_\Phi V_{\rm ne} + J(\rho_\phi) + E_{\rm xc}(\rho_\Phi).$$

The Kohn-Sham approach is *in principle* exact in the sense that if (3.10) is exactly satisfied, then $E_0^{\rm KS} = E_0$ for all molecular systems. On the other hand, as the exact exchange-correlation functional is not known explicitly, it must be approximated in practice to perform numerical calculations. For this reason, there is not one, but a whole zoology of Kohn-Sham models, corresponding to different approximations of $E_{\rm xc}(\rho)$. The simplest Kohn-Sham model actually used in practice is obtained using the *Local Density Approximation* (LDA) introduced by Kohn and Sham [54] (see also [79]). The resulting model is mathematically very similar to the so-called Xα model [84] where the exchange-correlation functional is approximated by the Dirac local exchange term:

$$E_{\rm xc}^{\rm X\alpha}(\rho) = -C_D\int_{\mathbb{R}^3} \rho^{4/3},$$

where $C_D = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}$ is the Dirac constant. Other more refined exchange-correlation functionals have been developed in the past 30 years, leading to the *Generalized Gradient Approximation* (GGA) [78], meta-GGA functionals [88], hybrid functionals [6], range-separated functionals [92] ... The Kohn-Sham ground state energy and density can be obtained by solving the Kohn-Sham equations (deduced from the Euler-Lagrange equations

associated with (3.11) by a simple algebraic manipulation)

$$\begin{cases} \rho^0(x) = \displaystyle\sum_{i=1}^{N} |\phi_i(x)|^2 \\ \mathfrak{h}_{\rho^0}\,\phi_i = \varepsilon_i \phi_i, \qquad \varepsilon_1 < \varepsilon_2 \leq \varepsilon_3 \leq \cdots \\ \displaystyle\int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \\ \mathfrak{h}_{\rho^0} = -\dfrac{1}{2}\Delta + V^{\mathrm{H}}_{\rho^0} + V^{\mathrm{xc}}_{\rho^0} \\ -\Delta V^{\mathrm{H}}_{\rho^0} = 4\pi \left(\rho^0 - \displaystyle\sum_{k=1}^{M} z_k \delta_{\mathbf{R}_k} \right), \end{cases} \tag{3.12}$$

where the Kohn-Sham potential $V^{\mathrm{KS}}_{\rho^0} = V^{\mathrm{H}}_{\rho^0} + V^{\mathrm{xc}}_{\rho^0}$ is the sum of the Hartree potential $V^{\mathrm{H}}_{\rho^0} = \rho_0 \star |\cdot|^{-1}$ and of the exchange-correlation potential $V^{\mathrm{xc}}_{\rho^0} = \frac{\partial E_{\mathrm{xc}}}{\partial \rho}(\rho^0)$. The differences between the various Kohn-Sham models lay in the form of the exchange-correlation potential. For the Xα model, we have $V^{\mathrm{xc,X}\alpha}_{\rho} = -\frac{4}{3} C_{\mathrm{D}} \rho^{1/3}$. As for Hartree-Fock, the Kohn-Sham equations have the mathematical form of a nonlinear elliptic eigenvalue problem. The Kohn-Sham model, as well as its generalizations (extended Kohn-Sham model, spin-polarized Kohn-Sham model) and approximations (reduced Hartree-Fock model), have been studied mathematically in [3, 45, 58, 64, 85]. A interesting related topic concerning the construction of density functionals for the Coulomb interaction is addressed in [30].

3.2.3. Quantum Monte Carlo methods. Quantum Monte Carlo (QMC) methods (see e.g. [4] and references therein) aim at solving the N-body quantum problem by means of stochastic algorithms. The term QMC encompasses several classes of methods including variational Monte Carlo (VMC), diffusion Monte Carlo (DMC), and path integral Monte Carlo. In the framework of electronic structure calculation, the most commonly used are VMC and DMC methods. For convenience, we use in this section the shorthand notation $\mathbf{x} = (x_1, \cdots, x_N)$,

$$V(\mathbf{x}) = -\sum_{i=1}^{N} \sum_{k=1}^{M} \frac{z_k}{|x_i - \mathbf{R}_k|} + \sum_{1 \leq i < j \leq N} \frac{1}{|x_i - x_j|} \quad \text{and} \quad H_N = -\frac{1}{2}\Delta + V,$$

where Δ is the Laplace operator in the $3N$-dimensional space.

VMC methods allow one to efficiently compute the energy of a large class of (non necessarily normalized) electronic wavefunctions ψ. The name *Variational* Monte Carlo originates from the fact that this approach can be used, in the spirit of usual variational methods, to seek an approximation of the ground state energy E_0 (and of a ground state wavefunction ψ_0) by minimizing the Rayleigh quotient $\frac{\langle \psi | H_N | \psi \rangle}{\langle \psi | \psi \rangle}$ over a family $\{\psi^{\mathbf{p}},\ \mathbf{p} \in \mathcal{P}\}$ of trial wave functions depending on a set of parameters $\mathbf{p}$. VMC methods are based on the observation that

$$\frac{\langle \psi | H_N | \psi \rangle}{\langle \psi | \psi \rangle} = \frac{\displaystyle\int_{\mathbb{R}^{3N}} E^{\psi}_{\mathrm{L}}(\mathbf{x})\, |\psi(\mathbf{x})|^2\, d\mathbf{x}}{\displaystyle\int_{\mathbb{R}^{3N}} |\psi(\mathbf{x})|^2\, d\mathbf{x}},$$

where the scalar field $E^{\psi}_{\mathrm{L}}(\mathbf{x}) = [H_N \psi](\mathbf{x})/\psi(\mathbf{x}) = -\frac{1}{2}\frac{\Delta\psi(\mathbf{x})}{\psi(\mathbf{x})} + V(\mathbf{x})$ is called the local

energy. Hence,

$$\frac{\langle\psi|H_N|\psi\rangle}{\langle\psi|\psi\rangle} = \lim_{\mathcal{N}\to\infty} \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} E_{\mathrm{L}}^{\psi}(\mathbf{X}^n), \tag{3.13}$$

where $(\mathbf{X}^n)_{n\geq 1}$ are points of $\mathbb{R}^{3N}$ sampling the probability distribution

$$d\mu_\psi(\mathbf{x}) = \frac{|\psi(\mathbf{x})|^2}{\int_{\mathbb{R}^{3N}} |\psi|^2}\, d\mathbf{x}.$$

The VMC method consists in approximating the right-side of (3.13) by an empirical mean for large, but finite values of $\mathcal{N}$. Note that if ψ is an eigenfunction of H_N associated with the eigenvalue E, $E_{\mathrm{L}}^{\psi}(\mathbf{x}) = E$ almost everywhere, so that the variance of E_{L}^{ψ} vanishes. In this extreme case, the relation

$$\frac{\langle\psi|H_N|\psi\rangle}{\langle\psi|\psi\rangle} = \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} E_{\mathrm{L}}^{\psi}(\mathbf{X}^n)$$

in fact holds true whatever $\mathcal{N}$ and the realizations $(\mathbf{X}^n)_{1\leq n\leq \mathcal{N}}$ of the random variable with law μ_ψ. Most often, VMC calculations are performed with trial wavefunctions ψ that are good approximations of a ground state wavefunction ψ_0. Consequently, $E_{\mathrm{L}}^{\psi}(\mathbf{x})$ usually is a function of low variance (with respect to the probability distribution μ_ψ). This is the reason why, in practice, the empirical mean $\frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}} E_{\mathrm{L}}^{\psi}(\mathbf{X}^n)$ is a fairly accurate approximation of $\frac{\langle\psi|H_N|\psi\rangle}{\langle\psi|\psi\rangle}$, even for relatively "small" values of $\mathcal{N}$.

Of course, the quality of this approximation depends on the way the points $(\mathbf{X}^n)_{n\geq 1}$ are generated. The standard sampling method currently used for VMC calculations is a Metropolis-Hastings algorithm based on a biased random walk in the configuration space $\mathbb{R}^{3N}$ [4].

Let us now turn to the DMC method [4]. For the sake of simplicity, we assume that the ground state energy E_0 is a *simple* eigenvalue of H_N, considered as an operator on $\mathcal{H}_N$, and we denote by $g = E_1 - E_0$ the spectral gap between the ground state energy E_0 and the first excited state energy E_1. The DMC method is based on the following observation. Let $\psi_I \in \mathcal{D}_N$. The unique solution $\psi(t,\mathbf{x})$ in $C^0(\mathbb{R}_+, \mathcal{D}_N) \cap C^1(\mathbb{R}_+, \mathcal{H}_N)$ of the parabolic problem

$$\left\{ \begin{array}{l} \dfrac{\partial\psi}{\partial t}(t,\mathbf{x}) = -(H_N\psi(t,\cdot))(\mathbf{x}) = \dfrac{1}{2}\Delta\psi(t,\mathbf{x}) - V(\mathbf{x})\psi(t,\mathbf{x}), \\ \psi(0,\mathbf{x}) = \psi_I(\mathbf{x}), \end{array} \right. \tag{3.14}$$

reads $\psi(t,\cdot) = e^{-tH_N}\psi_I$ and is such that

$$\|\exp(E_0 t)\ \psi(t) - \langle\psi_0|\psi_I\rangle\ \psi_0\|_{L^2} \leq \|\psi_I - \langle\psi_0|\psi_I\rangle\ \psi_0\|_{L^2}\ \exp(-gt),$$

where as above, ψ_0 denotes an L^2-normalized ground state of H_N. If moreover $\langle\psi_0|\psi_I\rangle \neq 0$, one also has

$$0 \leq E(t) - E_0 \leq \frac{\langle\psi_I|H|\psi_I\rangle - E_0}{|\langle\psi_0|\psi_I\rangle|^2}\, e^{-gt}, \quad \text{where} \quad E(t) = \frac{\langle\psi_I|H_N|\psi(t)\rangle}{\langle\psi_I|\psi(t)\rangle}. \tag{3.15}$$

As equation (3.14) is posed on $\mathbb{R}^{3N}$, and as in addition, V has singularities, it seems difficult to numerically solve (3.14) with deterministic methods.

On the other hand, a stochastic representation formula of the solution to (3.14) is provided by the Feynman-Kac formula

$$\psi(t,\mathbf{x}) = \mathbb{E}\left(\psi_I\left(\mathbf{x}+\mathbf{W}_t\right)\,\exp\left(-\int_0^t V\left(\mathbf{x}+\mathbf{W}_s\right)\,ds\right)\right), \tag{3.16}$$

where the expectation $\mathbb{E}$ is over the $\mathbb{R}^{3N}$-valued Wiener process $(\mathbf{W}_t)_{t\geq 0}$, and could *a priori* be used to estimate E_0 [74]. As such, (3.16) is however not adapted to numerical simulations: it has indeed been observed that for a given $\mathbf{x}\in\mathbb{R}^{3N}$, the variance of the random variable

$$Y_t^{\mathbf{x}} = \psi_I\left(\mathbf{x}+\mathbf{W}_t\right)\,\exp\left(-\int_0^t V\left(\mathbf{x}+\mathbf{W}_s\right)\,ds\right)$$

increases very quickly with time.

In practice, one makes use of an importance sampling technique. If the importance function ψ_I the DMC method is based upon is well-chosen, the ground state energy is approximated with a very good accuracy. In most cases, taking for ψ_I a Hartree-Fock ground state is sufficient to recover 90% of the correlation energy (the correlation energy is defined as the difference between the exact ground state energy E_0 and the energy E_0^{HF} of the Hartree-Fock ground state); for molecular systems in which the main part of the correlation energy is non-dynamical, that is when the ground state ψ_0 is badly approximated by a single Slater determinant, but fairly well approximated by a linear combination of a few Slater determinants, it is however necessary to consider multi-configurational importance functions [47]. The DMC method works as follows. Assume that the importance function ψ_I is continuous and such that the fields

$$\mathbf{b}^{\psi_I}(\mathbf{x}) = \frac{\nabla\psi_I(\mathbf{x})}{\psi_I(\mathbf{x})} \quad \text{and} \quad E_{\mathrm{L}}^{\psi_I}(\mathbf{x}) = \frac{(H_N\psi_I)(\mathbf{x})}{\psi_I(\mathbf{x})} = -\frac{1}{2}\frac{\Delta\psi_I(\mathbf{x})}{\psi_I(\mathbf{x})} + V(\mathbf{x})$$

exist for almost every $\mathbf{x}\in\mathbb{R}^{3N}$ and can be calculated with a reasonable computational cost (for instance, $\mathbf{b}^{\psi_I}(\mathbf{x})$ and $E_{\mathrm{L}}^{\psi_I}(\mathbf{x})$ can be computed in $O(N^4)$ operations if ψ_I is a Slater determinant). Consider the function

$$f_1(t,\mathbf{x}) = \psi_I(\mathbf{x})\,\psi(t,\mathbf{x}), \tag{3.17}$$

where ψ is the solution of (3.14). The energy $E(t)$ defined by (3.15) also reads

$$E(t) = \frac{\displaystyle\int_{\mathbb{R}^{3N}} E_{\mathrm{L}}^{\psi_I}(\mathbf{x})\,f_1(t,\mathbf{x})\,d\mathbf{x}}{\displaystyle\int_{\mathbb{R}^{3N}} f_1(t,\mathbf{x})\,d\mathbf{x}},$$

and an elementary calculation shows that f_1 is solution to the equation

$$\frac{\partial f}{\partial t} = \frac{1}{2}\Delta f - \mathrm{div}\left(\mathbf{b}^{\psi_I} f\right) - E_{\mathrm{L}}^{\psi_I} f, \qquad f(0,\mathbf{x}) = |\psi_I(\mathbf{x})|^2. \tag{3.18}$$

The above partial differential equation can be interpreted as the Fokker-Planck equation of a drift-diffusion process with source term. This leads us to considering the stochastic process defined by the stochastic differential equation (SDE)

$$d\mathbf{X}_t^{\mathbf{x}} = \mathbf{b}^{\psi_I}\left(\mathbf{X}_t^{\mathbf{x}}\right)\,dt + d\mathbf{W}_t, \quad \mathbf{X}_0^{\mathbf{x}} = \mathbf{x}, \tag{3.19}$$

the function

$$f_2(t,\mathbf{x}) = |\psi_I(\mathbf{x})|^2\, \mathbb{E}\left(\exp\left(-\int_0^t E_{\rm L}^{\psi_I}(\mathbf{X}_s^{\mathbf{x}})ds\right)\right), \tag{3.20}$$

and the real-valued function of time

$$E^{\rm DMC}(t) = \frac{\mathbb{E}\left(E_{\rm L}^{\psi_I}(\mathbf{X}_t)\exp\left(-\int_0^t E_{\rm L}^{\psi_I}(\mathbf{X}_s)ds\right)\right)}{\mathbb{E}\left(\exp\left(-\int_0^t E_{\rm L}^{\psi_I}(\mathbf{X}_s)ds\right)\right)}. \tag{3.21}$$

If the field $\mathbf{b}^{\psi_I}$ were regular enough and well-behaved at infinity (globally Lipschitz for instance), the SDE (3.19) would be well-posed by classical results (see e.g. [87]). Under the additional condition that the function $E_{\rm L}^{\psi_I}$ is bounded below, the functions f_1 and f_2 respectively defined by (3.17) and (3.20), would coincide, as well as the two quantities of interest $E(t)$ and $E^{\rm DMC}(t)$ defined by (3.15) and (3.21). This ideal scenario is encountered in the simulation of bosons, where the function ψ_I can be chosen positive everywhere, regular enough, and well-behaved at infinity. The situation is more delicate for fermions, as the field $\mathbf{b}^{\psi_I} = \frac{\nabla\psi_I}{\psi_I}$ is singular on the nodal surfaces of ψ_I. Under some technical assumptions we do not spell out in detail here, which are fulfilled for toy models (a system of non-interacting fermions confined in a harmonic potential), but should probably be refined to fully cover the case of electrons interacting with point nuclei, it is established in [20] that the SDE (3.19) has a unique solution, and that for all $\mathbf{x} \in U = \mathbb{R}^{3N} \setminus \psi_I^{-1}(0)$, the function $\mathbb{R}_+ \ni t \mapsto \mathbf{X}_t^{\mathbf{x}} \in \mathbb{R}^{3N}$ is in $C^0(\mathbb{R}_+, \mathcal{C}(\mathbf{x}))$, where $\mathcal{C}(\mathbf{x})$ is the connected component of U containing $\mathbf{x}$. In particular, the trajectories of (3.19) cannot cross the nodal surfaces $\psi_I^{-1}(0)$. This is due to the fact that close to the nodal surfaces, the random variable $\delta(t) = \mathrm{dist}(\mathbf{X}_t^{\mathbf{x}}, \psi_I^{-1}(0))$ behaves as the solutions to the SDE

$$dx_t = {x_t}^{-1}\, dt + dB_t,$$

where $(B_t)_{t\geq 0}$ is a one-dimensional Wiener process, which are known to stay away from zero almost surely in finite times. On the other hand, and similar to the case when $\mathbf{b}^{\psi_I}$ is globally Lipschitz, the random variable $\mathbf{X}_t^{\mathbf{x}}$ has a density $p(t,\mathbf{x},\mathbf{y})$ and the function $(\mathbf{x},\mathbf{y}) \mapsto \psi_I(\mathbf{x})^2 p(t,\mathbf{x},\mathbf{y})$ is symmetric. In the fermionic setting, the function f_2 defined by (3.20) still is a solution to (3.18) in the distributional sense, but it is not equal to f_1. More precisely, it holds

$$f_2(t,\mathbf{x}) = \psi_I(\mathbf{x})\phi(t,\mathbf{x}),$$

where $\phi(t,\mathbf{x})$ is the unique solution in $C^0(\mathbb{R}_+, \mathcal{D}_N) \cap C^1(\mathbb{R}_+, \mathcal{H}_N)$ to

$$\left\{\begin{array}{l} \dfrac{\partial\phi}{\partial t}(t,\mathbf{x}) = \dfrac{1}{2}\Delta\phi(t,\mathbf{x}) - V(\mathbf{x})\phi(t,\mathbf{x}), \\ \phi(0,\mathbf{x}) = \psi_I(\mathbf{x}), \\ \phi(t,\mathbf{x}) = 0 \text{ on } \psi_I^{-1}(0). \end{array}\right. \tag{3.22}$$

Problem (3.22) differs from problem (3.14) through the additional homogeneous Dirichlet condition that ϕ vanishes on the nodal surfaces of ψ_I. As a consequence, $E^{\rm DMC}(t)$ differs from $E(t)$ and it holds [20]

$$\lim_{t\to+\infty} E^{\rm DMC}(t) = E_0^{\rm DMC},$$

where

$$E_0^{\mathrm{DMC}} = \inf \left\{ \langle \psi | H_N | \psi \rangle,\ \psi \in \mathcal{Q}_N,\ \|\psi\|_{L^2} = 1,\ \psi = 0 \text{ on } \psi_I^{-1}(0) \right\}.$$

Obviously $E_0^{\mathrm{DMC}} \geq E_0$, and the equality holds if and only if the nodal surfaces ofψ_I coincide with those of a ground state ψ_0 of H_N. The systematic bias introduced by the choice of a function ψ_I which does not have the same nodes as ψ_0 (which is the case in practice), is called the *fixed node error*. Getting rid of the fixed node error in quantum Monte Carlo simulations of fermions is one of the major challenges in computational physics.

4. Some recent advances and open questions

4.1. Towards certified numerical methods. As we have seen, the Hartree-Fock and Kohn-Sham models are infinite-dimensional constrained (non-convex) optimization problems, whose Euler-Lagrange equations are nonlinear elliptic eigenvalue problems. Computing numerically the Hartree-Fock or Kohn-Sham ground states therefore requires

1. a discretization method to transform the infinite-dimensional problem into a finite-dimensional one;
2. an iterative algorithm to solve the so-obtained finite-dimensional contrained optimization problem, or the associated Euler-Lagrange equations.

Theoretical chemists and computational physicists have devoted a lot of effort to the development of efficient discretization methods for electronic structure calculation. The most common discretization methods are Gaussian atomic orbitals [49] for molecules, and planewaves [36] for solid state physics and materials science. New approaches based on wavelets [43] or discontinuous Galerkin methods [71] have also been recently introduced. We refer to [13] for a recent review of the iterative algorithms for solving the discretized Hartree-Fock and Kohn-Sham problems. The development of fast numerical methods is an active field of research [8, 33, 34, 42, 52, 69, 70, 72]. Note that optimized black-box algorithms are not yet available, and that very few convergence results for existing algorithms [22, 23, 60] have been established so far.

One of the challenges for the next decade is to construct accurate and robust error estimators with respect to the various numerical parameters used to perform the calculation (truncation of the discretization basis in variational approximations, number of points in numerical quadratures, convergence thresholds for iterative algorithms, ...). These estimators could then be used to adapt in real time the numerical parameters to equilibrate the various sources of error. As a result, the maximal committed error could be certified and the simulation time strongly reduced in comparison with the current usual approach consisting in fixing *a priori* the numerical parameters and testing the quality of the chosen parameters by checking that the results hardly change when refining the discretization and reducing the convergence thresholds. Important progress in this direction has been made in the last few years; in particular, optimal *a priori* error bounds on nonlinear eigenvalue problems have recently been obtained [14], and these results have been applied to the analysis of Kohn-Sham models [15, 29]. The construction of *a posteriori* error estimators for Kohn-Sham is a current active field of research.

4.2. Multiscale models. An important, very challenging, mathematical and numerical problem is concerned with the coupling of quantum chemistry models with coarser models in view of simulating larger molecular systems. Such approaches include QM/MM models [90] (for the development of which Karplus, Levitt and Warshel were awarded the 2013 Nobel Prize in Chemistry), and implicit solvation models [91]. QM/MM models consist in cutting the molecular system (say a drug interacting with a protein) into two subsystems (the drug and the active site of the protein on the one hand, the rest of the protein on the other hand) and in treating the first one (small but key) with quantum mechanics (QM) and the second one (large and playing the role of the "environment") with classical molecular mechanics (MM). Understanding such coupling between quantum and classical models is an essentially open mathematical question.

So far, we assumed that the molecule under study could be considered as an isolated system, which is almost never the case in practice. In particular, most chemical reactions take place in the liquid phase. In principle, we could apply the models previously introduced to a "supermolecule" consisting of the solute molecule and a big number of solvent molecules. This is however not doable in practice for two reasons. First, this would dramatically increase the size of the system; second, we would need to properly sample and average over the configurations of the solvent molecules, which is very difficult and most often unfeasible in practice. Implicit solvation models, which date back to Born, Kirkwood and Onsager, consist in replacing all the solvent molecules but the few ones strongly interacting with the solute, with an effective continuous medium accounting for long-range electrostatics. This amounts to replacing the Poisson equation in (3.12) by the inhomogeneous elliptic equation

$$-\mathrm{div}\left(\epsilon\nabla V_{\rho^0}^{\mathrm{H}}\right) = 4\pi\left(\rho^0 - \sum_{k=1}^{M} z_k \delta_{\mathbf{R}_k}\right), \tag{4.1}$$

with $\epsilon(x) = 1$ inside a cavity Ω containing the molecule (see Fig. 4.1) and $\epsilon(x) = \epsilon_{\mathrm{s}}$ outside Ω, where ϵ_{s} is the macroscopic dielectric permittivity of the solvent (about 80 for water). A numerical method coupling Schwarz's domain decomposition method with integral equations has recently been proposed to solve the so-called COSMO approximation of (4.1) in the framework of classical and quantum molecular models [75]. This allows one to perform geometry optimization on large molecules in solution with a limited extra-cost with respect to the same calculation *in vacuo*.

Implicit solvation models are widely used in chemistry and give satisfactory results in many cases. On the other hand, they fail in other cases, in particular in the presence of strong interactions between the solute and the solvent. Also, the definition of the molecular cavity Ω is a touchy business, and some physical properties may strongly depend on the chosen definition. For all these reasons, constructing better solvation models using mathematical tools such as model reduction techniques is an interesting problem of major importance for applications.

4.3. Thermodynamic limits and the crystal problem. In contrast with the contents of the previous two subsections, we now present a purely theoretical problem. Consider a cluster with L^3 identical atoms, put the L^3 nuclei on the sites of a cubic lattice (for simplicity) and compute the electronic ground state for this nuclear configuration (Fig. 4.1). Several questions are in order: when L goes to infinity,

(i) does the ground state energy per atom converge?

(ii) does the ground state electronic density converge?

(iii) does it converge to a periodic density?

(iv) can those quantities be obtained by solving a periodic problem on a unit cell?

(v) if nuclei are allowed to relax to their equilibrium positions for finite L, do we obtain a periodic crystal in the limit?

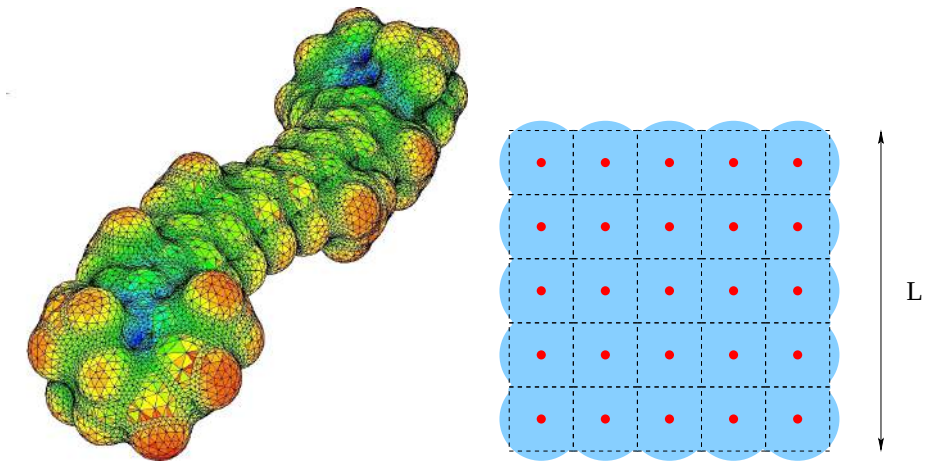


Figure 4.1. Left: a molecular cavity for the carotene molecule used in implicit salvation models (courtesy P. Laug, Inria). Right: a cubic cluster of size L. The thermodynamic limit problem consists in studying the limit of the electronic structure of the cluster when L goes to infinity.

The first four questions fall into the scope of thermodynamic limit problems. The fifth issue is called the crystal problem. The thermodynamic limit problems in the terms stated above have been the subject of many outstanding contributions in the context of various energy models and various physical systems [26, 38, 48, 65, 66, 81]. The thermodynamic limit problem is completely solved for the TF [67] and TFW [26] models. For Hartree-Fock and Kohn-Sham type models, the fundamental issues (ii)-(iii) remain open. The two key difficulties are first that the latter models are not convex (convexity plays a crucial role in the analysis of the TFW model) and second that the number of molecular orbitals ϕ_i to be dealt with, or equivalently the rank of the one-body density matrix $\gamma_\Phi = \sum_i |\phi_i\rangle\langle\phi_i|$, are also growing to infinity (in the TFW model, only the density ρ is relevant). For the HF model, partial results have been established [27, 44]. The results are partial in the sense that it is needed to *postulate*, in addition to the periodicity of the set of nuclei, the periodicity of the ground state density matrix in the limit. A simplified version of the HF model, namely the reduced Hartree-Fock model however allows for a complete proof [27].

4.4. Crystals with defects and disordered systems. In the previous section, we dealt with perfect crystals, while the really interesting systems for the applications are crystals with defects. The difficulty is that such systems contain infinitely many interacting electrons and have no symmetry allowing us to reduce the problem to a periodic cell, as is the case for perfect crystals. In the past few years, some progress has been made in the theoretical understanding of Kohn-Sham models for insulating and semiconducting crystals with local

defects [17, 18, 24], but the cases of metals and extended defects is still open (see [40] for the case of a uniform electron gas and [19] for the TFW model). On the numerical side, very little is known: in particular, there is no completely satisfactory method to deal with charged defects in insulators and semi-conductors.

A huge amount of literature has been devoted to modeling electrons in random materials. In most cases, electrons are considered as non-interacting particles subjected to a stationary random empirical potential $V(\omega, x)$ [25]. The analysis of the electronic properties of the material then reduces to the analysis of the spectral properties of the associated random Schrödinger operator $H(\omega) = -\frac{1}{2}\Delta + V(\omega, \cdot)$ acting on $L^2(\mathbb{R}^3)$. A remarkable property of random Schrödinger operators is that, under some ergodicity and integrability assumptions on V, the spectrum of $H(\omega)$ is deterministic: there exists a closed set $\Sigma \in \mathbb{R}$ such that $\sigma(H(\omega)) = \Sigma$ almost surely. Similar results hold for the density of states (that is, losely speaking, the number of quantum states per unit volume) of the Hamiltonian $H(\omega)$ [11]. Interesting questions are concerned with the nature of the spectrum (point spectrum, absolutely continuous spectrum, ...), which is related to the electronic transport properties of the material. We refer to [25, 86] and references therein for more details on the linear case.

Serious additional difficulties arise for models with interacting electrons. Recent results have been obtained on the thermodynamic limit of disordered quantum systems composed of interacting particles with short-range (Yukawa) interactions [9, 21, 93]. The case of long-range (Coulomb) interactions was investigated in [10]. The existence of a thermodynamic limit is proven, but the limit is not identified. An interesting open problem consists in studying the case of rare but possibly large random perturbations, which corresponds to the physical situation of doped semiconductors with low concentration of impurities. This question has been successfully addressed in [53] in the case of a linear model of non-interacting electrons. The case of interacting electrons with short-range interactions is dealt with in [56]. The problem is still open for electrons in Coulomb interactions.

Acknowledgements. I am grateful to V. Ehrlacher, G. Gontier, M. Luskin and G. Stoltz for useful comments on this manuscript.

The bibliography mostly contains *mathematical* contributions to the field of electronic structure calculation. Due to the lack of space, very few of the hundreds of relevant references of the physics and chemistry literatures are cited.

References

[1] Ambrosio, L., Figalli, A., Friesecke, G., Giannoulis, J., and Paul, T., *Semiclassical limit of quantum dynamics with rough potentials and well posedness of transport equations with measure initial data*, Comm. Pure Appl. Math. **64** (2011), 1199–1242.

[2] Ambrosio, L., Friesecke, G., and Giannoulis, J., *Passage from quantum to classical molecular dynamics in the presence of Coulomb interactions*, CPDE **35** (2010), 1490–1515.

[3] Anantharaman A., *Cancès, E., Existence of minimizers for Kohn-Sham models in quantum chemistry*, Ann. I. H. Poincaré, **26** (2009), 2425–2455.

[4] Bach, V. and Delle Site, L. (Eds.), *Many-electron approaches in physics, chemistry and mathematics. A Multidisciplinary View*. Springer, 2014.

[5] Bach, V., Lieb, E.H., Loss, M., and Solovej, J.P., *There are no unfilled shells in unrestricted Hartree-Fock theory*, Phys. Rev. Lett. **72** (1994), 2981–2983.

[6] Becke, A.D., *Density-functional thermochemistry. III. The role of exact exchange*, J. Chem. Phys. **98** (1993), 5648–5652.

[7] Benedikt, U., Auer, H., Espig, M., Hackbusch, W., and Auer, A.A., *Tensor representation techniques in post-Hartree-Fock methods : matrix product state tensor format*, Mol. Phys. **111** (2013), 2398–2413.

[8] Benzi, M., Boito, P., and Razouk, N., *Decay properties of spectral projectors with applications to electronic structure*, SIAM Rev. **55** (2013), 3–64.

[9] Blanc, X., Le Bris, C., and Lions, P.-L., *On the energy of some microscopic stochastic lattices, Part I*, Arch. Rat. Mech. Anal. **184** (2007), 303–340.

[10] Blanc, X. and Lewin, M., *Existence of the thermodynamic limit for disordered quantum Coulomb systems*, J. Math. Phys. **53** (2012), 095209.

[11] Bourgain, J. and Klein, A., *Bounds on the density of states for Schrödinger operators*, Invent. Math. **194** (2013), 41–72.

[12] Bourquin, R., Gradinaru, V., and Hagedorn, G.A., *Non-adiabatic transitions near avoided crossings: theory and numerics*, J. Math. Chem. **50** (2012), 602–619.

[13] Cancès, E., *Self-consistent field (SCF) algorithms*, in: Encyclopedia of Applied and Computational Mathematics, B. Engquist (Ed.), Springer, to appear.

[14] Cancès, E., Chakir, R., and Maday, Y., *Numerical analysis of nonlinear eigenvalue problems*, J. Sci. Comput. **45** (2010), 90–117.

[15] _____, *Numerical analysis of the planewave discretization of orbital-free and Kohn-Sham models*, M2AN **46** (2012), 341–388.

[16] Cancès, E., Defranceschi, M., Kutzelnigg, W., Le Bris, C., and Maday, Y., *Computational quantum chemistry: A primer*, in Handbook of Numerical Analysis, Vol. X: Computational chemistry, eds. P. Ciarlet, C. Le Bris, North-Holland, 2003, 3–270.

[17] Cancès, E., Deleurence, A., and Lewin, M., *A new approach to the modelling of local defects in crystals: the reduced Hartree-Fock case*, Commun. Math. Phys. **281** (2008), 129–177.

[18] _____, *Non-perturbative embedding of local defects in crystalline materials*, J. Phys.: Condens. Mat. **20** (2008), 294213.

[19] Cancès, E. and Ehrlacher, V., *Local defects are always neutral in the Thomas-Fermi-von Weiszäcker theory of crystals*, Arch. Ration. Mech. Anal. **202** (2011), 933–973.

[20] Cancès, E., Jourdain, B., and Lelièvre, T., *Quantum Monte Carlo simulation of fermions. A mathematical analysis of the fixed-node approximation*, M3AS **16** (2006), 1403–1440.

[21] Cancès, E., Lahbabi, S., and Lewin, M., *Mean-field models for disordered crystals*, J. Math. Pures Appl. **100** (2013), 241–274.

[22] Cancès, E. and Le Bris, C., *On the convergence of SCF algorithms for the Hartree-Fock equations*, M2AN **34** (2000), 749–774.

[23] ———, *Can we outperform the DIIS approach for electronic structure calculations?*, Int. J. Quantum Chem. **79** (2000), 82–90.

[24] Cancès, E. and Lewin, M., *The dielectric permittivity of crystals in the reduced Hartree-Fock approximation*, Arch. Ration. Mech. Anal. **197** (2010), 139–177.

[25] Carmona, R. and Lacroix, J., *Spectral theory of random Schrödinger operators*, Birkhäuser, 1990.

[26] Catto, I., Le Bris, C., and Lions, P.-L., *Mathematical theory of thermodynamic limits: Thomas-Fermi type models*, Oxford University Press, 1998.

[27] ———, *On the thermodynamic limit for Hartree-Fock type models*, Ann. I. H. Poincaré **18** (2001), 687–760.

[28] Ceperley, D.M. , *Path integrals in the theory of condensed helium*, Rev. Mod. Phys. **67** (1995), 279–355.

[29] Chen, H., Gong, X., He, L, Yang, Z., and Zhou, A., *Numerical analysis of finite dimensional approximations of Kohn-Sham models*, Adv. Comput. Math. **38** (2013), 225–256.

[30] Cotar, C., Friesecke, G., and Klüppelberg, C., *Density functional theory and optimal transportation with Coulomb cost*, Comm. Pure Appl. Math. **66** (2013), 548–599.

[31] Dirac, P.A.M., *Quantum Mechanics of Many-Electron Systems*, Proc. Royal Soc. London Ser. A **123** (1929), 714–733.

[32] Dreizler, R., Gross, E.K.U, *Density functional theory*, Springer Verlag, 1990.

[33] E, W. and Lu, J., *The Kohn-Sham equation for deformed crystals*, Mem. Amer. Math. Soc. **221** no. 1040 (2013).

[34] E, W., Li, T., and Lu, J., *Localized basis of eigen-subspaces and operator compression*, Proc. Natl. Acad. Sci. **107** (2010), 1273.

[35] Esteban, M.J., Lewin, M., and Séré, E., *Variational methods in relativistic quantum mechanics*, Bull. Amer. Math. Soc. **45** (2008), 535–593.

[36] Fiolhais, C. Nogueira, F., and Marques, M.A.L., (Eds.), *A Primer in Density Functional Theory*, Lecture Notes in Physics, Vol. **620**, Springer, 2003.

[37] Flad, H.J., Hackbusch, W., Khoromskij, B.N., and Schneider, R., *Concepts of data-sparse tensor-product approximation in many-particle modelling*, In: Matrix methods: theory, algorithms and applications, World Scientific, 2010, 313–347.

[38] Fefferman, C., *The thermodynamic limit for a crystal*, Commun. Math. Phys. **98** (1985), 289–311.

[39] Fermanian Kammerer, C., Gérard, P., and Lasser, C., *Wigner measure propagation and Lipschitz conical singularity for general initial data*, Arch. Ration. Mech. Anal. **209** (2013), 209–236.

[40] Frank, R., Lewin, M., Lieb, E.H., and Seiringer, R., *Energy cost to make a hole in the Fermi sea*, Phys. Rev. Lett. **106** (2011), 150402.

[41] Friesecke, G., *The multiconfiguration equations for atoms and molecules: charge quantization and existence of solutions*, Arch. Rat. Mech. Anal. **169** (2003), 35–71.

[42] García-Cervera, C.J., Lu, J., Xuan, Y., and E, W., *A linear scaling subspace iteration algorithm with optimally localized non-orthogonal wave functions for Kohn-Sham density functional theory*, Phys. Rev. B **79** (2009), 115110.

[43] Genovese, L.,Videau, B., Ospici, M., Deutsch, T., Goedecker, S, and Méhaut, J.-F., *Daubechies wavelets for high performance electronic structure calculations: The BigDFT project*, C. R. Mec. **339** (2011), 149–164.

[44] Ghimenti, M. and Lewin, M., *Properties of periodic Hartree-Fock minimizers*, Calculus of Variations and Partial Differential Equations **35** (2009), 39–56.

[45] Gontier, D., *N-Representability in noncollinear spin-polarized density functional theory*, Phys. Rev. Lett. **111** (2013), 153001.

[46] Griesemer, M. and Hantsch, F., *Unique solutions to Hartree-Fock equations for closed shell atoms*, Arch. Ration. Mech. Anal. **203** (2012), 883–900.

[47] Grossman, G.C., *Benchmark quantum Monte Carlo calculations*, J. Chem. Phys. **117** (2002), 1434–1440.

[48] Hainzl, C., Lewin, M., and Solovej, J.P., *The thermodynamic limit of quantum Coulomb systems*, Adv. Math. **221** (2009), 454–546.

[49] Helgaker, T., Jorgensen, P., and Olsen, J., *Molecular electronic-structure theory*, Wiley, 2000.

[50] Hohenberg, P. and Kohn, W., *Inhomogeneous electron gas*, Phys. Rev. **136** (1964), B864-B871.

[51] Kato, T., *Pertubation theory for linear operators*, Springer-Verlag, 1995.

[52] Khoromkaia, V., Khoromskij, B.N., and Schneider, R., *QTT representation of the Hartree and exchange operators in electronic structure calculations*, Comp. Meth. in Applied Math. **11** (2011), 327–341.

[53] Klopp, F., *An asymptotic expansion for the density of states of a random Schrödinger operator with Bernoulli disorder*, Random Oper. Stoch. Eq. **3** (1995), 315–331.

[54] Kohn, W. and Sham, L.J., *Self-consistent equations including exchange and correlation effects*, Phys. Rev. **140** (1965), A1133.

[55] Korobov, V., and Yelkhovsky, A., *Ionization potential of the helium atom*, Phys. Rev. Lett. **87** (2001), 193003.

[56] Lahbabi, S., *The reduced Hartree-Fock model for short-range quantum crystals with nonlocal defects*, Ann. H. Poincaré, to appear.

[57] Le Bris, C., *A general approach for multiconfiguration methods in quantum molecular chemistry*, Ann. Inst. H. Poincaré 11 (1994), 441–484.

[58] _____, *Quelques problèmes mathématiques en chimie quantique moléculaire*, PhD thesis, Ecole Polytechnique, 1993.

[59] Lelièvre, T., Rousset, M., and Stoltz, G., *Free energy computations: a mathematical perspective*, Imperial College Press, 2010.

[60] Levitt, A., *Convergence of gradient-based algorithms for the Hartree-Fock equations*, M2AN **46** (2012), 1321–1336.

[61] Levy, M., *Universal variational functionals of electron densities, first order density matrices, and natural spin-orbitals and solution of the V-representability problem*, Proc. Natl. Acad. Sci. USA **76** (1979), 6062-6065.

[62] Lewin, M., *Solutions of the multiconfiguration equations in quantum chemistry*, Arch. Rat. Mech. Anal. **171** (2004), 83–114.

[63] Lieb, E.H., *Thomas-Fermi and related theories of atoms and molecules*, Rev. Mod. Phys. **53** (1981), 603–641.

[64] _____, *Density Functional for Coulomb systems*, Int. J. Quantum Chem. **24** (1983), 243–277.

[65] Lieb, E.H. and Lebowitz, J.L., *The constitution of matter: existence of thermodynamics for systems composed of electrons and nuclei*, Adv. Math. **9** (1972), 316–398.

[66] Lieb, E.H. and Seiringer, R., *The stability of matter in quantum mechanics*, Cambridge University Press, 2010.

[67] Lieb, E.H. and Simon, B., *The Thomas-Fermi theory of atoms, molecules and solids*, Adv. Math. **23** (1977), 22–116.

[68] _____, *The Hartree-Fock theory for Coulomb systems*, Comm. Math. Phys. **53** (1977), 185–194.

[69] Lin, L., M. Chen, Yang, C., and He, L., *Accelerating atomic orbital-based electronic structure calculation via pole expansion and selected inversion*, J. Phys.: Condens. Matter **25** (2013) 295501.

[70] Lin, L., Lu, J., Ying, L., Car, R., and E. W., *Fast algorithm for extracting the diagonal of the inverse matrix with application to the electronic structure analysis of metallic systems*, Comm. Math. Sci. **7** (2009), 755–777.

[71] Lin, L., Lu, J., L. Ying, and E. W., *Adaptive local basis set for Kohn-Sham density functional theory in a discontinuous Galerkin framework I*, Total energy calculation, J. Comp. Phys. **231** (2012), 2140–2154.

[72] Lin, L. and Yang, C., *Elliptic preconditioned for accelerating the self-consistent field iteration in Kohn-Sham density functional theory*, SIAM J. Sci. Comp. **35** (2013), S277–S298.

[73] Lions, P.-L., *Solutions of Hartree-Fock equations for Coulomb systems*, Comm. Math. Phys. **109** (1987), 33–97.

[74] _____, *Remarks on mathematical modelling in quantum chemistry. Computational Methods in Applied Sciences*, Wiley, New York 1996, pp. 22–23.

[75] Lipparini, F., Lagardère, L., Scalmani, G., Stamm, B, Cancès, E., Maday, Y., Piquemal, J.-P., Frisch, M., and Mennucci, B., *Quantum calculations in solution for large to very large molecules: a new linear scaling QM/continuum approach*, J. Chem. Phys. Lett., in press.

[76] Lu, J. and F. Otto, *Nonexistence of minimizer for Thomas-Fermi-Dirac-von Weizsacker model*, Comm. Pure Appl. Math., to appear.

[77] Panati, G., Spohn, H., and Teufel, S., *The time-dependent Born-Oppenheimer approximation*, M2AN **41** (2007), 297–314.

[78] Perdew, J.P., Burke, K., and Ernzerhof, M., *Generalized gradient approximation made simple*, Phys. Rev. Lett. **77** (1996), 3865–3868.

[79] Perdew, J.P. and Zunger, A., *Self-interaction correction to density-functional approximations for many-electron systems*, Phys. Rev. B **23** (1981), 5048–5079.

[80] Radhakrishnan, B. and Gavini, V., *Effect of cell size on the energetics of vacancies in aluminum studied via orbital-free density functional theory*, Phys. Rev. B **82** (2010), 094117.

[81] Ruelle, D., *Statistical Mechanics. Rigorous results*, Imperial College Press and World Scientific Publishing, 1999.

[82] Schneider, R., *Analysis of the projected coupled cluster method in electronic structure calculation*, Numer. Math. **113** (2009), 433–471.

[83] Sinha, D., Maitra, R., and Mukherjee, D., *Generalized antisymmetric ordered products, generalized normal ordered products, ordered and ordinary cumulants and their use in many electron correlation problem*, Comput. Theor. Chem. **100** (2013), 62–70.

[84] Slater, J.C., *A simplification of the Hartree-Fock method*, Phys. Rev. **81** (1951), 385–390.

[85] Solovej, J.P., *Proof of the ionization conjecture in a reduced Hartree-Fock model*, Invent. Math. **104** (1991), 291–311.

[86] Stollmann, P., *Caught by disorder: bound states in random media*, Birkhäuser, 2001.

[87] Stroock D.W. and Varadhan S.R.S., *Multidimensional diffusion processes*, Classics in Mathematics 233. Springer, Berlin, 2006.

[88] Tao, J.M., Perdew, J.P., Staroverov, V.N., and Scuseria, G.E., *Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids*, Phys. Rev. Lett. **91** (2003), 146401.

[89] Teller, E., *On the stability of molecules in the Thomas-Fermi theory*, Rev. Mod. Phys. **34** (1962), 627–631.

[90] Thiel, W. and Hummer, G., *Nobel 2013 Chemistry: Methods for computational chemistry*, Nature **504**, 96–97.

[91] Tomasi, J., Mennucci, B., and Cammi, R., *Quantum Mechanical Continuum Solvation Models*, Chem. Rev. **105** (2005), 2999.

[92] Toulouse, J., Colonna, F., and Savin, A., *Long-range/short-range separation of the electron-electron interaction in density-functional theory*, Phys. Rev. A **70** (2005), 062505.

[93] Veniaminov, N.A., *The existence of the thermodynamic limit for the system of interacting quantum particles in random media*, Ann. H. Poincaré **14** (2013), 64.

[94] Wang, Y.A., Govind, N., and Carter, E.A., *Orbital-free kinetic-energy density functionals with a density-dependent kernel*, Phys. Rev. B **60** (1999), 16350–16358.

[95] Zhislin, G.M. and Sigalov, A.G., *The spectrum of the energy operator for atoms with fixed nuclei on subspaces corresponding to irreducible representations of the group of permutations*, Izv. Akad. Nauk SSSR Ser. Mat. **29** (1965), 835–860.

Université Paris Est, Ecole des Ponts and Inria, 6 & 8 avenue Blaise Pascal, 77455 Marne-la-Vallée, France

E-mail: cances@cermics.enpc.fr

Sparse analysis

Anna C. Gilbert

Abstract. The goal of this lecture is to give you an introduction to the mathematics, algorithms, and applications in the field of sparse analysis, including sparse approximation and compressive sensing. Both of these problems contain a wealth of challenging algorithmic problems, novel uses of existing mathematical techniques, as well as mathematical innovations. Coupled with these theoretical challenges are practical engineering questions that both support and motivate the mathematical innovations. The fundamental mathematical problem is that of solving an under-determined linear system. Despite learning in high school algebra that such problems are "impossible" to solve, mathematicians, computer scientists, and engineers attempt to do so in a myriad of fields, applications, and settings. This problem arises in signal and image compression, theoretical computer science, algorithms for massive, streaming data sets, high-throughput biological screens, and in the design of analog-to-digital converters.

Mathematics Subject Classification (2010). Primary 42; Secondary 68.

Keywords. Sparse approximation, compressive sensing, sublinear algorithms, compression, sparse signal recovery.

1. Introduction

The goal of this lecture is to give you an introduction to the mathematics, algorithms, and applications in the field of sparse analysis, including sparse approximation and compressive sensing. Despite learning in high school mathematics that solving an *under-determined* linear system of the form $Ax = b$ is impossible, mathematicians, computer scientists, and engineers attempt to do so in a myriad of fields, applications, and settings. This problem arises in signal and image compression where A plays the role of a redundant dictionary, b is the image or signal to compress, and x is the collection of "transform" coefficients that captures the exact linear combination of dictionary elements which closely approximate b. In this case, solving $Ax = b$ amounts to finding a sparse or parsimonious linear representation over A that is close to the input signal b. In theoretical computer science, and streaming algorithms in particular, the matrix A is generated at random from a particular distribution, x accumulates the stream of updates to a vector or distribution of values, and b is the "sketch" of that stream of updates. Given A and b, a streaming algorithm strives to compute quickly statistical or combinatorial information about x. In these applications, the number of rows of the matrix A is smaller than the number of columns, sometimes exponentially so. In compressive sensing, one *designs* a matrix A, takes linear measurements of a signal using A, $Ax = b$, and from these measurements and knowledge of A, one tries to reconstruct information about x.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

One mathematical feature that allows us to solve many variants of this under-determined linear system problem is *sparsity*; roughly speaking, as long as the vector x does not contain too many non-zero components (or has a few dominating components), we can "solve" the under-determined system $Ax = b$. Common variants of the problem are collectively called *sparse approximation* (SA) problems.

We begin our study of sparse signal analysis with a simple example of image compression. In Figure 1.1 we take an image and compute its 2d wavelet transform. This transform is an orthogonal transform so the transform produces a set of wavelet coefficients that are nothing other than a change of basis from the original image. The left panel of Figure 1.1 shows both the original image and its 2d wavelet transform. Notice that many of the wavelet coefficients in the lower left figure are dark or close to zero. If we were to quantize all of the coefficient values, we would spend a considerable fraction of our total bits quantizing the small coefficient values. Instead, let's spend most of our bits on the large (in absolute value) coefficients and set to zero those coefficients that are smaller (in absolute value) than a threshold θ. If we then compute the inverse wavelet transform of the thresholded coefficients, we obtain a close approximation to the original image (top right panel). Those few coefficients we retained capture almost all of the important features of the original image and there are considerably fewer of these coefficients than pixels in the original image so we have produced a lossy compressed representation that consists of a few non-zero wavelet coefficients.

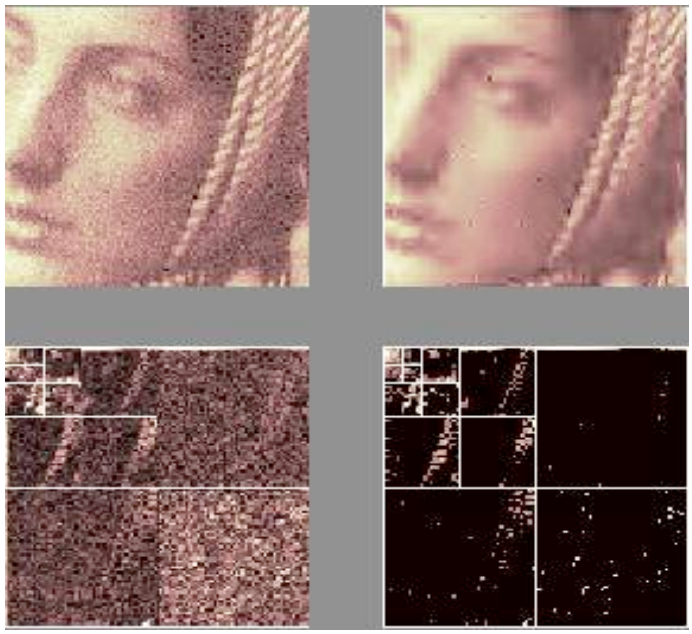


Figure 1.1. original image (top left); 2D wavelet transform of image (bottom left); threshold wavelet coefficients (bottom right); inverse wavelet transform of thresholded wavelet coefficients (top right)

Let's make this discussion a bit more formal. Let Φ denote orthogonal matrix that corresponds to our orthonormal basis; the vectors of the basis are the columns of the matrix Φ. And let x denote the signal that we wish to compress. The encoding algorithm in Algorithm 1 that we used above is simple.

Algorithm 1: Encoding algorithm that computes the coefficients of an image over an orthonormal basis and then thresholds them.

Input: Φ and x
Output: t sparse vector
$c = \Phi^* x$ // compute orthogonal transform
$t = \Theta(c)$ // threshold small coefficients

The decoding (or reconstruction) algorithm that we use to synthesize the image from its compressed representation is also straightforward and is shown in Algorithm 2.

Algorithm 2: Reconstruction or synthesis algorithm that we use to reconstruct an image from its compressed representation.

Input: Φ and t
Output: x
$\hat{x} = \Phi t$

This encoding algorithm is an example of a nonlinear encoding algorithm and a linear decoding algorithm. The encoding algorithm is nonlinear in the sense that *which* coefficients are retained in the thresholding process depend on the input signal x.

This configuration is a typical one for modern codecs; more computational resources are spent encoding signals or images so that the decoder can be relatively simple. Later in this course, we will see that compressive sensing attempts to invert this typical allocation of resources with a *linear* encoder and a *nonlinear* decoder.

Nonlinear encoding with a fixed orthonormal basis Φ for signal compression has several advantages and one large disadvantage. Both the encoding and decoding algorithms are relatively easy to compute and, given a signal x, the compression that the encoder produces is the best set of coefficients for *that* signal; i.e., the compression is *instance optimal*. It is important to note that the compressibility of the input signal x depends on the fixed orthonormal basis Φ and that the compression is only as good as the basis Φ is for representing x. Given a large class of input signals x, it is challenging to design a single orthonormal basis that compresses all of the input signals well. This is especially true if the class consists of natural images. If we fix a single input x and then design one orthonormal basis Φ, it is easy to produce a basis that compresses x perfectly (simply take the first vector in the basis to be $x \, \|x\|_2$ and all the other vector orthogonal to x); however, the orthonormal basis we construct may not compress many other signals!

Throughout this discussion, we have referred informally to the terms *sparsity* and *compressibility* somewhat interchangeably but it's important that we have precise definitions of these terms before we proceed.

Definition 1.1. The *sparsity* of a vector $x \in \mathbb{R}^d$ or $\mathbb{C}^d$ is the number of non-zero entries in x. We denote this quantity by $\|x\|_0$ despite the fact that this is not a proper norm.

Definition 1.2. Suppose that we sort in decreasing absolute value the entries of a vector $x \in \mathbb{R}^d$. The rate of decay of these entries is the *compressibility* of x. Let $|x_{(\ell)}|$ be the sorted vector x, then the compressibility is the exponent α in the relationship

$$|x_{(\ell)}| \sim \ell^{-\alpha}.$$

See Figure 1.2 for an example of a compressible signal and what effect thresholding entries in a compressible vector has on its sparsity.

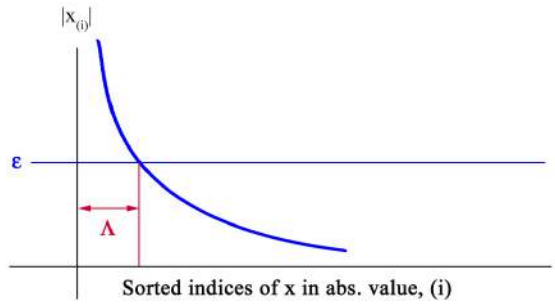


Figure 1.2. The set Λ consists of those entries of a vector x that are larger (in absolute value) than a threshold ϵ.

These definitions assume that the basis in which we represent or compress the signal x is the canonical basis but, if instead we use a different orthonormal basis Φ, the vector in question is the vector of coefficients Φx. Sparsity or compressibility are two ways of measuring the number of items we need to encode to represent a signal; the fewer the items, the fewer the bits, and the greater the compression. On a more philosophical level, sparsity or compressibility is a mathematical version of Occam's razor; we strive to capture the essential features of a signal.

2. Complete and redundant dictionaries

One of the biggest drawbacks of using a single orthonormal basis for compression, especially for a large class of input signals, is that the compression is only as good as the basis is "right" for that class and it is difficult to design a single orthonormal basis that compresses well a large class of signals. It is much easier, instead, to use more than orthonormal basis from which to represent a single input signal x or a class of signals. With apologies to George Orwell,

> if one orthonormal basis is good, surely two (or more) are better...especially for images.

Figure 2.1 shows the different features that are present in an image, from the brush-like texture of the mandrill's hair to the pointillist-like residual that remains after we remove a linear combination of both a few wavelets and a few wavelet packets. No single orthonormal basis is rich enough (or large enough!) to include vectors that resemble all of these features.

Which bring us to a rigorous definition of a dictionary for signal representation.

Definition 2.1. A *dictionary* Φ in $\mathbb{R}^d$ (or $\mathbb{C}^d$) is a collection $\{\varphi_\ell\}_{\ell=1}^N \subset \mathbb{R}^d$ (or $\mathbb{C}^d$) of unit-norm vectors: $\|\varphi_\ell\|_2 = 1$ for all ℓ.

- Elements are called *atoms*
- If span$\{\varphi_\ell\} = \mathbb{R}^d$, the dictionary is *complete*

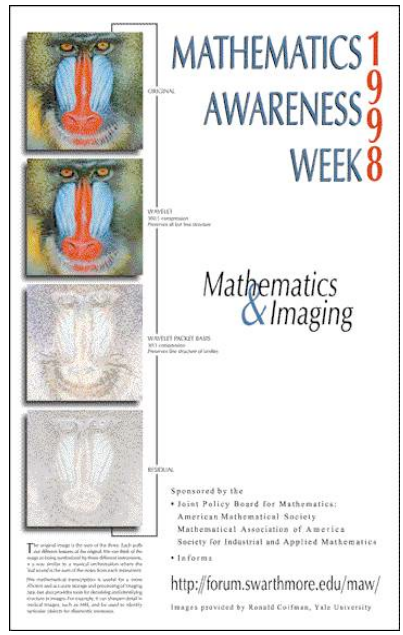


Figure 2.1. Different bases capture different features of an image.

- If $\{\varphi_\ell\}$ are linearly dependent and the dictionary is complete, the dictionary is *redundant*.

We represent a dictionary as a matrix

$$\Phi = \begin{bmatrix} \varphi_1 & \varphi_2 & \dots & \varphi_N \end{bmatrix}$$

with atoms as the columns of the matrix so that

$$\Phi c = \sum_\ell c_\ell \varphi_\ell.$$

See Figure 2.2 for the various linear algebraic operations.

2.1. Examples.

(1) **Fourier-Dirac:** $\Phi = [\,\mathcal{F}\,|\,I\,]$. In $\mathbb{C}^d$, we define the dictionary as:

$$\varphi_\ell(t) = \frac{1}{\sqrt{d}} \mathrm{e}^{2\pi \mathrm{i} \ell t/d} \quad \ell = 1, 2, \dots, d$$
$$\varphi_\ell(t) = \delta_\ell(t) \quad \ell = d+1, d+2, \dots, 2d.$$

It is of size $2d$ and the atoms are shown in Figure 2.3.

(2) **DCT-Wavelets:** $\Phi = [\,\mathcal{F}\,|\,\mathcal{W}\,]$. In $\mathbb{R}^{d\times d}$ (or two-dimensional images), we have the union of the discrete cosine transform (DCT) (a basis traditionally used to compress

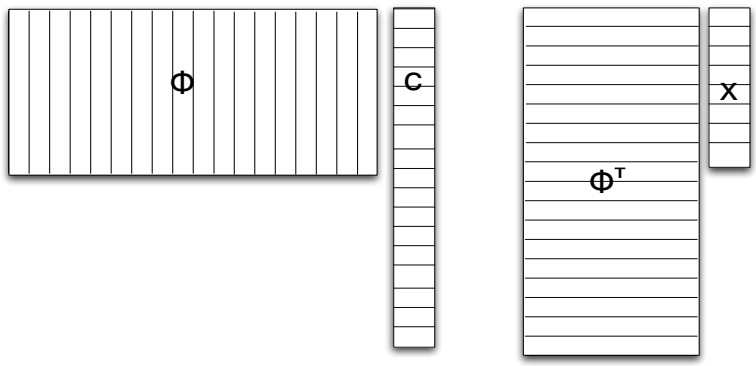


Figure 2.2. The operation Φc produces a linear combination of dictionary atoms and is used to *synthesize* a signal in $\mathbb{R}^d$ (left) while the operation $\Phi^T x$ produces a vector of dot products of x with the dictionary atoms or the *analysis* of a signal over a dictionary (right).

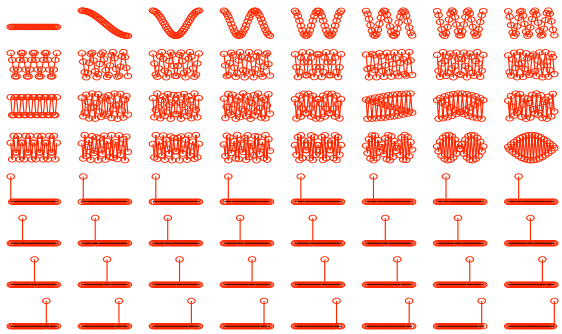


Figure 2.3. The first four rows are the Fourier basis and the second four rows show the Dirac basis for $d = 32$.

video frames) and two-dimensional wavelets (another plausible basis for compressing images). This dictionary is of size $2d^2$, or twice as large as $d \times d$ images themselves. See Figure 2.4 for several representative atoms in this dictionary.

(3) **Wavelet packets:** Φ in $\mathbb{R}^d$ consists of $d \log d$ different wavelet packets which can be arranged into $2^{O(d)}$ different orthonormal bases. See Figure 2.5 for an example.

3. Sparse approximation

If we use the convention that we represent a signal as a linear combination over a redundant dictionary Φc, then there are two competing metrics for the quality of this representation: (i) the cost of the representation itself (typically, how sparse or how compressible c is) and (ii) the error in the approximation $x \approx \Phi c$. In signal or image compression, we usually trade-off the *rate* of the encoding with the *distortion* that it produces as compared to the

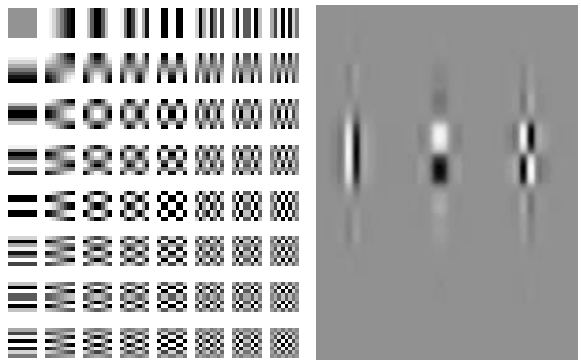


Figure 2.4. The DCT of an image (left) and three different 2-dimensional wavelets (right).

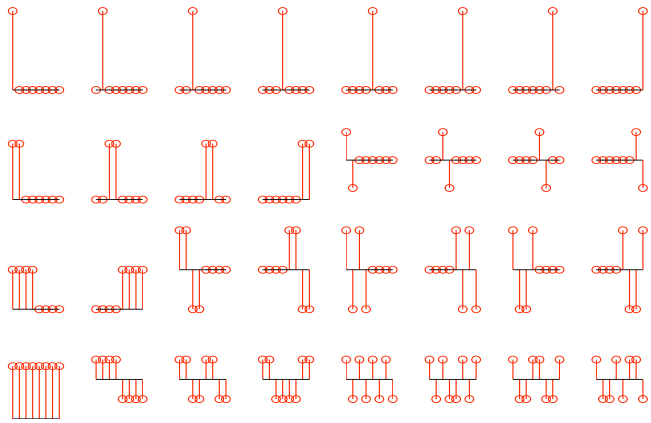


Figure 2.5. The Haar wavelet packet dictionary for $d = 8$.

original signal. We will define these terms slightly differently from that of typical source coding; nonetheless, our definitions address this inherent trade-off in how expensive versus how accurate the representation or compression is.

3.1. Basic problems. The following three problems capture precisely three types of compression or signal representation: lossless compression, lossy compression that meets an error (or distortion) tolerance, and lossy compression that meets a memory (or rate) tolerance.

- EXACT. Given a vector $x \in \mathbb{R}^d$ and a complete dictionary Φ, solve

$$\arg\min_c \|c\|_0 \quad \text{s.t.} \quad x = \Phi c$$

  i.e., find a sparsest representation of x over Φ.

- ERROR. Given $\epsilon \geq 0$, solve

$$\arg\min_c \|c\|_0 \quad \text{s.t.} \quad \|x - \Phi c\|_2 \leq \epsilon$$

i.e., find a sparsest approximation of x that achieves error ϵ.

- SPARSE. Given $k \geq 1$, solve

$$\arg\min_c \|x - \Phi c\|_2 \quad \text{s.t.} \quad \|c\|_0 \leq k$$

i.e., find the best approximation of x using k atoms.

Now that we have formalized three fundamental sparse approximation problems, there are two directions to pursue: (i) how to compute sparse approximations and (ii) if we could compute them, how accurate would they be (or how should we evaluate the output of the algorithms)?

Theorem 3.1. *EXACT is NP-complete.*

Corollary 3.2. *All of the other sparse approximation problems SPARSE, ERROR, EXACT are all NP-hard.*

This result is terribly pessimistic. It says that given any polynomial time algorithm for any of the three sparse approximation problems, there is a dictionary Φ and a signal x for which the algorithm returns an incorrect answer. But we may not see the worst case combination Φ and x in many practical applications. Furthermore, the types of redundant dictionaries we use in signal or image compression are fairly natural or structured and are not combinations of arbitrary vectors. It is also important to note that this hardness result depends on both the dictionary and input vector being arbitrary. There are other types of problem instances for which the sparse approximation problems are not hard. Other types of problem instances include fixed redundant dictionaries, random dictionaries, random signals, fixed signals, etc. Whether each instance admits a feasible solution depends on the type of instance, on the choice of distribution (in the case of random dictionaries or random signals), and on the choice of fixed dictionary or signal. A tremendous amount of research in the last 10 years has centered around these different instance types, including, or especially, that of compressive sensing. In order to understand the compressive sensing results, it's worthwhile to understand where it fits in to the bigger picture of sparse approximation. Figure 3.1 summarizes these different instances.

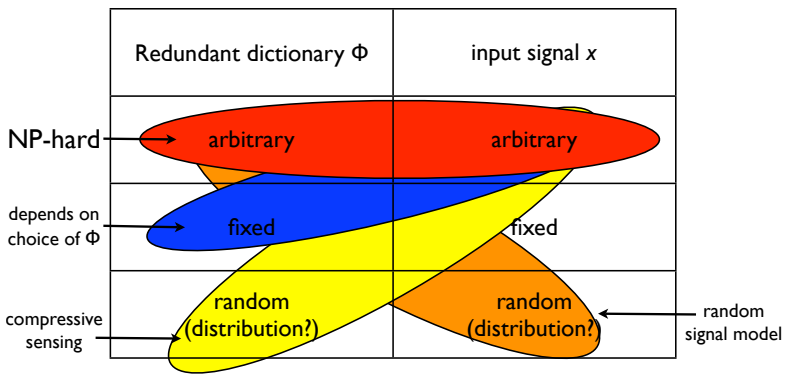


Figure 3.1. The hardness of the sparse approximation problems depends on the type of problem instance.

3.2. A greedy iterative algorithm for SPARSE. We will focus our attention on a particular greedy, iterative algorithm for this discussion. By greedy, iterative algorithms we mean those that choose the best atom (or collection thereof) at each step of the algorithm. The algorithm that we analyze is Orthogonal Matching Pursuit, shown in Algorithm 3. This algorithm is essentially a prototype greedy algorithm but it differs in one important place; the method by which we calculate the coefficients c_λ that participate in the representation is more complicated. At each step, the coefficients are the solution to a least-squares problem. That is, we compute the orthogonal projection of x onto the subspace spanned by the columns indexed by Λ, the active set of atoms in the representation. The orthogonal projection step is precisely where OMP gets the leading "O" in its name. Because at each step of the algorithm, the representation is the orthogonal projection of x onto the span of the selected atoms, each selected atom in the representation appears only once. That is, the residual is orthogonal to the representation so atoms that already participate in the representation cannot be selected again at a later step of the algorithm. Many greedy algorithms have a similar outline and there are number of adjustments one could make to this basic architecture.

Algorithm 3: Orthogonal Matching Pursuit (OMP).

Input: Φ, x, and k
Output: c
Initialize $a = 0$, $\Lambda = \emptyset$
for $j = 1$ **to** k **do**
  $r = x - a$
  $\lambda_j = \arg\max_\ell |\langle r, \varphi_\ell \rangle|$
  $\Lambda = \Lambda \cup \{\lambda_j\}$
  Find c_λ for $\lambda \in \Lambda$ that minimizes $\|x - \sum_{\lambda \in \Lambda} c_\lambda \varphi_\lambda\|_2$
  Update $a = \sum_{\lambda \in \Lambda} c_\lambda \varphi_\lambda$
end

We would like to know when OMP can correctly identify sparse representations and do so in a reasonable number of steps. Suppose x has k-sparse representation

$$x = \sum_{\lambda \in \Lambda} c_\ell \varphi_\lambda$$

with $|\Lambda| = k$. That is, the optimal representation c_{OPT} is non-zero only on the set Λ. For OMP to find the optimal representation c_{OPT}, it's sufficient for OMP to identify the support set Λ (once we have found Λ, we compute the coefficients in c_Λ by least-squares and this is the closest approximation in ℓ_2.)

To that end, define

$$\Phi_\Lambda = \begin{bmatrix} \varphi_{\lambda_1} & \varphi_{\lambda_2} & \cdots & \varphi_{\lambda_k} \end{bmatrix}_{\lambda_s \in \Lambda} \quad \text{and}$$
$$\Psi_\Lambda = \begin{bmatrix} \varphi_{\lambda_1} & \varphi_{\lambda_2} & \cdots & \varphi_{\lambda_{N-k}} \end{bmatrix}_{\lambda_s \notin \Lambda}$$

Define *greedy selection ratio* $\rho(r)$ which is a function of the residual vector r

$$\rho(r) = \frac{\max_{\lambda \notin \Lambda} |\langle r, \varphi_\lambda \rangle|}{\max_{\lambda \in \Lambda} |\langle r, \varphi_\lambda \rangle|} = \frac{\|\Psi_\Lambda^T r\|_\infty}{\|\Phi_\Lambda^T r\|_\infty};$$

this is the ratio of the largest dot product between the residual and atoms not in Λ to the largest dot product between the residual and atoms in Λ. We claim that OMP chooses good atoms in Λ at each iteration if and only if $\rho(r_t) < 1$ for the residual r_t in each iteration $t = 1, \ldots, k$. Note that we do not require that the smallest "good" dot product $|\langle r, \varphi_\lambda|$ with $\lambda \in \Lambda$ be greater than the largest "bad" dot product $|\langle r, \varphi_\lambda|$ with $\lambda \notin \Lambda$ as OMP selects the atom (in all of the dictionary Φ) corresponding to the *largest* dot product. We can now state a *sufficient* condition for OMP to identify Λ. This is known as the **Exact Recovery Condition** (ERC). See [10] for more information on this algorithm and its performance for sparse approximation problems.

Theorem 3.3. *A sufficient condition for OMP Λ after k steps is that*

$$\max_{\ell \notin \Lambda} \|\Phi_\Lambda^+ \varphi_\ell\|_1 < 1.$$

To clarify the geometric interpretation of this result, recall he pseudo-inverse of a matrix A is denoted $A^+ = (A^T A)^{-1} A^T$. The vector $A^+ x$ is a vector of coefficients that synthesizes best approximation of x using the atoms in the columns of A. The projector $P = AA^+$ is the orthogonal projector produces this best approximation.

Proof. We will prove this by induction. Let us observe that at the beginning of the algorithm, our approximation $a_0 = 0$, so $r_0 = x \in \mathrm{range}(\Phi_\Lambda)$. Our inductive hypothesis is that at the beginning of each iteration, the residual $r_t = x - a_t \in \mathrm{range}(\Phi_\Lambda)$; i.e., at iteration $t + 1$ assume that the previously selected atoms are in the support set, $\lambda_1, \lambda_2, \ldots, \lambda_t \in \Lambda$, so that $a_t \in \mathrm{range}(\Phi_\Lambda)$ and $r_t = x - a_t \in \mathrm{range}(\Phi_\Lambda)$.

We express the (transpose symmetric) orthogonal projector onto $\mathrm{range}(\Phi_\Lambda)$ as $(\Phi_\Lambda^+)^T \Phi_\Lambda^T$, therefore

$$(\Phi_\Lambda^+)^T \Phi_\Lambda^T r_t = r_t.$$

Then we can bound the greedy selection ratio at this iteration as

$$\begin{aligned} \rho(r_t) &= \frac{\|\Psi_\Lambda^T r_t\|_\infty}{\|\Phi_\Lambda^T r_t\|_\infty} = \frac{\|\Psi_\Lambda^T (\Phi_\Lambda^+)^T \Phi_\Lambda^T r_t\|_\infty}{\|\Phi_\Lambda^T r_t\|_\infty} \\ &\leq \|\Psi_\Lambda^T (\Phi_\Lambda^+)^T\|_\infty \\ &= \|\Phi_\Lambda^+ \Psi_\Lambda\|_1 \\ &= \max_{\ell \notin \Lambda} \|\Phi_\Lambda^T \varphi_\ell\|_1 < 1 \end{aligned}$$

Then OMP selects an atom from Λ at iteration t and since it chooses a new atom at each iteration, after k iterations, it has chosen all the atoms from Λ. □

This result is not particularly useful in that we cannot evaluate the ERC if we do not know Λ a priori. Instead, we rely on a proxy, a geometric condition on the dictionary that we can check, regardless of the input signal.

Definition 3.4. The *coherence* of a dictionary is the largest dot product between distinct pairs of atoms

$$\mu = \max_{j \neq \ell} |\langle \varphi_j, \varphi_\ell \rangle|.$$

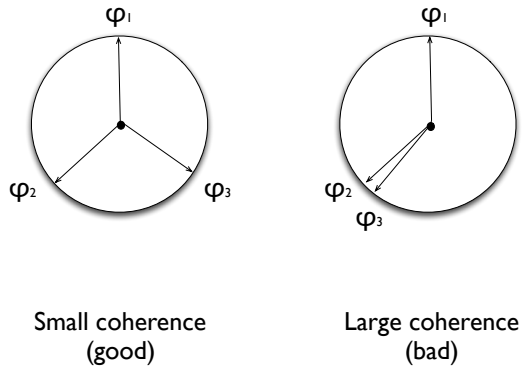


Figure 3.2. Coherence of two different collections of atoms in $\mathbb{R}^2$.

Figure 3.2 illustrates two different collections of atoms in $\mathbb{R}^2$; one with small coherence (which is a good property) and one with large coherence (which is a bad property).

One might wonder just how small the coherence of a collection of N atoms can be in dimension d, especially if N is quite large compared to d. Fortunately, there are a number of results of this type from coding theory (as redundant dictionaries are nothing other than spherical codes). The most straightforward lower bound on the coherence is known as the Welch bound.

Theorem 3.5. *For a $d \times N$ redundant dictionary,*

$$\mu \geq \sqrt{\frac{N-d}{d(N-1)}} \approx \frac{1}{\sqrt{d}}.$$

It is always useful to have at hand several examples of large, incoherent dictionaries from which to draw intuition.

- Fourier–Dirac, $N = 2d$, $\mu = \frac{1}{\sqrt{d}}$
- wavelet packets, $N = d \log d$, $\mu = \frac{1}{\sqrt{2}}$
- There are large dictionaries with coherence close to the lower (Welch) bound; e.g., Kerdock codes, $N = d^2$, $\mu = 1/\sqrt{d}$

Finally, we can state a result that is checkable and that does not depend on the input signa.

Theorem 3.6 (Tropp'04). *The ERC holds whenever $k < \frac{1}{2}(\mu^{-1} + 1)$. Therefore, OMP can recover any sufficiently sparse signals.*

The lower bound on coherence is $\mu \geq 1/\sqrt{d}$ so, for most redundant dictionaries, $k < C\sqrt{d}$.

Proof. Let us assume $|\Lambda| = k$. We expand the Gram matrix

$$\Phi_\Lambda^T \Phi_\Lambda = I + A$$

and bound $\|A\|_1 \leq (k-1)\mu$.

Now, let us bound the exact recovery expression in two pieces

$$\begin{aligned}\max_{\ell\notin\Lambda}\|\Phi_\Lambda^+\varphi_\ell\|_1 &= \max_{\ell\notin\Lambda}\|(\Phi_\Lambda^T\Phi_\Lambda)^{-1}\Phi_\Lambda^T\varphi_\ell\|_1\\ &\le \|(\Phi_\Lambda^T\Phi_\Lambda)^{-1}\|_1 \max_{\ell\notin\Lambda}\|\Phi_\Lambda^T\varphi_\ell\|_1.\end{aligned}$$

We estimate the second piece as:

$$\max_{\ell\notin\Lambda}\|\Phi_\Lambda^T\varphi_\ell\|_1 = \max_{\ell\notin\Lambda}\sum_{j\in\Lambda}|\langle\varphi_\ell,\varphi_j\rangle| \le k\mu$$

We estimate the first piece using our decomposition of the Gram matrix:

$$\begin{aligned}\|(\Phi_\Lambda^T\Phi_\Lambda)^{-1}\|_1 = \|(I+A)^{-1}\|_1 &= \|\sum_{j=0}^{\infty}(-A)^j\|_1\\ &\le \sum_{j=0}^{\infty}\|A\|_1^j = \frac{1}{1-\|A\|_1}\\ &\le \frac{1}{1-(k-1)\mu}\end{aligned}$$

It is sufficient that $k\mu + (k-1)\mu < 1$ or $k < \frac{1}{2}(\mu^{-1}+1)$ for the ERC to hold. □

4. Sparse signal recovery

There is one important variation of the sparse approximation problems that has captured the attention of mathematicians, computer scientists, engineers, and scientists alike. Indeed, tracking heavy hitters in high-volume, high-speed data streams [1], monitoring changes in data streams [4], designing pooling schemes for biological tests [7] (e.g., high throughput sequencing, testing for genetic markers), localizing sources in sensor networks [11, 12], and combinatorial pattern matching [2] are all quite different technological challenges, yet they can all be expressed in the same mathematical formulation. We have a signal $\mathbf{x}$ of length N that is sparse or highly compressible; i.e., it consists of k significant entries ("heavy hitters") which we denote by $\mathbf{x}_k$ while the rest of the entries are essentially negligible. We wish to acquire a small amount of information (commensurate with the sparsity) about this signal in a linear, non-adaptive fashion, and then use that information to quickly recover the significant entries. In a data stream setting, our signal is the distribution of items seen, while in biological group testing, the signal is proportional to the binding affinity of each drug compound (or the expression level of a gene in a particular organism). We want to recover the identities and values of only the heavy hitters which we denote by $\mathbf{x}_k$, as the rest of the signal is not of interest.

Mathematically[1], we have a signal $\mathbf{x}$ and an m-by-N measurement matrix $\mathbf{\Phi}$ with which we acquire measurements $\mathbf{y} = \mathbf{\Phi x}$, and, from these measurements $\mathbf{y}$, we wish to recover $\widehat{\mathbf{x}}$,

[1]The notation in this section will be slightly different from that in the previous section, to keep with the notation in the field. We hope that all readers are astute and can shift accordingly!

with $O(k)$ entries, such that

$$\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leq C \|\mathbf{x} - \mathbf{x}_k\|_2 .$$

Our goal is to design the measurement matrix $\mathbf{\Phi}$ and the decoding algorithm $\mathcal{A}$ in an optimal fashion. In this section, we discuss fundamental lower bounds for this problem as well as one particular sublinear algorithm that achieves these goals.

In the above applications, it is important both to take as few measurements as possible and to recover the heavy hitters extremely efficiently. Measurements correspond to physical resources (e.g., memory in data stream monitoring devices, number of screens in biological applications) and reducing the number of necessary measurements is critical these problems. In addition, these applications require efficient recovery of the heavy hitters—we test many biological compounds at once, we want to quickly identify the positions of entities in a sensor network, and we cannot afford to spend computation time proportional to the size of the distribution in a data stream application. In several of the applications, such as high throughput screening and other physical measurement systems, it is also important that the result be robust to the corruption of the measurements by an arbitrary noise vector ν_2. (It is less critical for digital measurement systems that monitor data streams in which measurement corruption is less likely.) For these reasons, we focus on "optimal" algorithms that are *sublinear* with respect to running time and take as few measurements as possible.

4.1. Fundamental results.

Definition 4.1. Define a measurement system $(\mathcal{A}, \Phi)_m$ to consist of an algorithm $\mathcal{A}$ and an $m \times N$ measurement matrix Φ. The algorithm $\mathcal{A}$ takes as input $y = \Phi x \in \mathbb{R}^m$ and Φ and returns $\widehat{x} \in \mathbb{R}^N$; i.e., $\mathcal{A}(\Phi x, \Phi) = \widehat{x}$.

We begin our discussion of signal recovery with (exactly) sparse vectors.

Definition 4.2. Define Σ_k to be the set

$$\{x \in \mathbb{R}^N \mid \|x\|_0 \leq k\};$$

i.e., the set of all (at most) k-sparse vectors in $\mathbb{R}^N$.

It is important that we watch the quantifiers in our specification so, for this portion of the discussion, we will insist on a measurement scheme $(\mathcal{A}, \Phi)_m$ with *one* single matrix Φ such that for all $x \in \Sigma_k$, the algorithm recovers x exactly, $\mathcal{A}(\Phi x, \Phi) = x$. Our first, most basic question is whether such a measurement system exists. Furthermore, with fixed N and k, what is the minimum number of rows m (i.e., minimum number of measurements) needed to recover all $x \in \Sigma_k$ exactly, regardless of the efficiency of the algorithm?

In order to answer these questions, we review some basic linear algebra. Recall the definition of the null space of a matrix.

Definition 4.3. The *nullspace* of a matrix Φ is given the set

$$\mathcal{N}(\Phi) = \left\{x \in \mathbb{R}^N \mid \Phi x = 0\right\}.$$

This is the set of vectors that Φ maps to zero.

There are several important properties of the null space that we will make use of:

(1) $\mathcal{N}(\Phi)$ is a subspace of $\mathbb{R}^N$,

(2) $0 \in \mathcal{N}(\Phi)$, and

(3) if Φ has rank r, then $\dim(\mathcal{N}(\Phi)) = N - r$.

The null space of Φ plays an important role in all of our sparse recovery results because without a priori conditions on x and Φ, we can recover x only up to the addition of elements in the null space of Φ as x and $x + z$ for $z \in \mathcal{N}(\Phi)$ have the same measurements $y = \Phi x = \Phi(x + z)$. To be more explicit, let us define for $\Phi x = y \in \mathbb{R}^m$

$$F(y) = \left\{ z \in \mathbb{R}^N \mid \Phi z = y \right\},$$

the set of all pre-images of y. Then, it is clear that $F(y) = x + \mathcal{N}(\Phi)$.

The following lemma (from [3]) provides a straightforward lower bound on the number of measurements m our measurement matrix must satisfy and what equivalent properties it must also possess in order to meet our requirements for a measurement scheme. Let $H \subseteq \{1, \cdots, N\}$ index a subset of columns of Φ and let Φ_H denote the $m \times |H|$ submatrix of Φ formed by the columns of Φ indexed by H.

Lemma 4.4. *If Φ is any $m \times N$ matrix and $2k \leq m$, then the following are equivalent:*

(1) *There is an algorithm $\mathcal{A}$ such that for all $x \in \Sigma_k$ $\mathcal{A}(\Phi x, \Phi) = x$.*

(2) $\Sigma_{2k} \cap \mathcal{N}(\Phi) = \{0\}$.

(3) *For any set H with $|H| = 2k$, the matrix Φ_H has rank $2k$.*

(4) *The symmetric matrix $(\Phi_H)^T (\Phi_H)$ is positive definite.*

Proof. Note that (2) $\implies$ (3) $\implies$ (4) by basic linear algebra, so we won't prove those implications.

(1) $\implies$ (2): We assume that there is an algorithm $\mathcal{A}$ such that $\mathcal{A}(\Phi x, \Phi) = x$,for all $x \in \Sigma_k$ and consider any $x \in \Sigma_{2k} \cap \mathcal{N}(\Phi)$. Our goal is to show that such a vector x must be zero. To that end, we decompose $x = x_1 - x_2$ with each portion $x_1, x_2 \in \Sigma_k$. Since x is in the null space of Φ, we know that:

$$\Phi x = 0 = \Phi x_1 - \Phi x_2$$

Thus,

$$\Phi x_1 = \Phi x_2.$$

Next, we apply $\mathcal{A}$ to both (equal) measurement vectors and we find that

$$\mathcal{A}(\Phi x_1, \Phi) = x_1 = x_2 = \mathcal{A}(\Phi x_2, \Phi)$$

and, hence, $x_1 = x_2$. Therefore, any vector x that is (at most) $2k$-sparse and in the null space of Φ must be the zero vector.

(2) $\implies$ (1): We assume that the only vector $x \in \Sigma_{2k} \cap \mathcal{N}(\Phi)$ is the zero vector. Our goal is to use this hypothesis to show the existence of an algorithm. To that end, we will *construct* an algorithm (ignoring its efficiency).

For any $y \in \mathbb{R}^m$, consider $F(y)$, the set of pre-images of y and define $\mathcal{A}(y, \Phi) = z$ to be $z \in F(y)$ with the smallest support; i.e.,

$$\mathcal{A}(y, \Phi) = \arg\min_{z \in F(y)} \|z\|_0.$$

Let $x_1 \in \Sigma_k$ and run $\mathcal{A}(\Phi x_1, \Phi)$. Suppose that the algorithm we have constructed fails to return the proper vector x_1 and that, instead, it returns $x_2 \neq x_1$ with $x_2 \in F(y)$ but $\|x_2\|_0 \leq \|x_1\|_0$. Clearly, $x_2 \in \Sigma_k$ (it has no more non-zero entries than x_1), $x_1 - x_2 \in \Sigma_{2k}$, and $\Phi(x_1 - x_2) = 0$, so $x_1 - x_2 \in \Sigma_{2k} \cap \mathcal{N}(\Phi)$. But our hypothesis tells us that $x_1 = x_2$. Hence, the algorithm constructed above must succeed. (The fact that we have demonstrated an algorithm that meets the requirements proves that one exists.) □

In fact, it is possible to construct matrices Φ with $m = 2k$ that possess the properties in the lemma. For any k and $N \geq 2k$, we can find a set Λ_N of N vectors in $\mathbb{R}^{2k}$ such that any $2k$ of them are linearly independent. For example, if $0 < x_1 < x_2 < \cdots < x_N$, then the matrix whose (i, j) entry is x_j^{i-1} satisfies these properties. Its $2k \times 2k$ minors are Vandermonde matrices which are non-singular. Unfortunately, these minors have large condition numbers when N is large and recovering $x \in \Sigma_k$ from $y = \Phi x$ is numerically unstable. Worse, these are (linear) algebraic properties of Φ and are somewhat limited as we rarely encounter *truly* sparse signals.

Unfortunately, most signals x are not *exactly* $k-$sparse, especially signals with some physical meaning attached to them. We would like to derive necessary and sufficient conditions on the measurement matrix Φ to guarantee the (approximate) recovery of more general signals. It is too much to ask that we recover the entire signal x; instead we recovery it up to the accuracy of the best k-term approximation to x, $\sigma_k(x)_p$. See Figure 4.1 for an illustration of this concept.

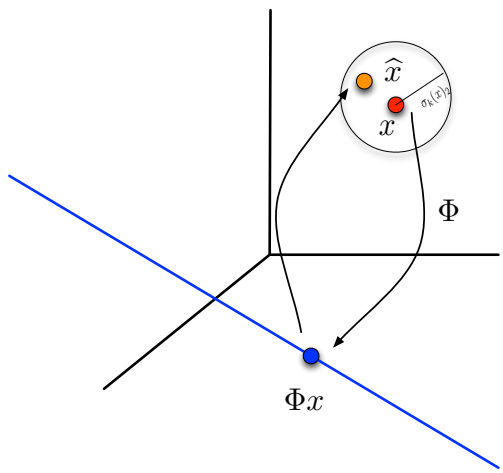


Figure 4.1. Recovery of x from Φx up to the accuracy of $\sigma_k(x)_2 = \|x - x_k\|_2$.

With this as our goal for signal recovery, we can state necessary and sufficient conditions on Φ for us to guarantee the existence of a recovery algorithm. The necessary condition, in particular, is referred to as *the null space condition*. See [3] for more information.

Theorem 4.5. *Given an $m \times N$ matrix Φ, a norm $\|\cdot\|_p$, and a sparsity value k, a **sufficient** condition for there to exist an algorithm $\mathcal{A}$ such that, for all x,*

$$\|x - \mathcal{A}(\Phi x, \Phi)\|_p \leq C_0 \|x - x_k\|_p$$

is that for all $n \in \mathcal{N}$

$$\|n\|_p \leq \frac{C_0}{2}\|n - n_{2k}\|_p. \tag{4.1}$$

A ***necessary*** *condition is that for all* $n \in \mathcal{N}$

$$\|n\|_p \leq C_0\|n - n_{2k}\|_p. \tag{4.2}$$

4.2. Sublinear algorithm. In this section, we discuss only at a high-level, a measurement system (i) that minimizes the number $m = O\big(k \log N/k\big)$ of measurements, (ii) for which the decoding algorithm $\mathcal{A}$ runs in *sublinear* time $O\big(k \log N/k\big)$, and (iii) the encoding and update times are optimal $O\big(N \log N/k\big)$ and $O\big(\log N/k\big)$, respectively. In order to achieve this, our algorithm is randomized; i.e., we specify a distribution on the measurement matrix $\mathbf{\Phi}$ and we guarantee that, for each signal, the algorithm recovers a good approximation with high probability over the choice of matrix. This result and considerably more detail can be found in [9]. This measurement system is essentially optimal. We summarize the results in Theorem 4.6.

Theorem 4.6. *There is an algorithm and distribution on matrices* $\mathbf{\Phi}$ *satisfying* $\max_{\mathbf{x}} \mathbb{E}[\|\mathbf{\Phi x}\|_2 / \|\mathbf{x}\|_2] = 1$ *such that, given* $\mathbf{\Phi x} + \nu_2$*, the parameters, and a concise description of* $\mathbf{\Phi}$*, the algorithm returns* $\widehat{\mathbf{x}}$ *with approximation error* $\|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2 \leq (1+\epsilon)\|\nu_1\|_2^2 + \epsilon\|\nu_2\|_2^2$ *with probability* $\frac{3}{4}$*. The algorithm runs in time* $k/\epsilon \log^{O(1)} N$ *and* $\mathbf{\Phi}$ *has* $O\big(k/\epsilon \log(N/k)\big)$ *rows. In expectation, there are* $O(\log^2(k)\log(N/k))$ *non-zeros in each column of* $\mathbf{\Phi}$*.*

The approximation $\widehat{\mathbf{x}}$ may have more than k terms. From [8], it is known that, if

$$\|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2 \leq (1+\epsilon^2)\|\mathbf{x} - \mathbf{x}_k\|_2^2 + \epsilon^2\|\nu_2\|_2^2,$$

then the truncation $\widehat{\mathbf{x}}_k$ of $\widehat{\mathbf{x}}$ to k terms satisfies

$$\|\mathbf{x} - \widehat{\mathbf{x}}_k\|_2^2 \leq (1+\Theta(\epsilon))\|\mathbf{x} - \mathbf{x}_k\|_2^2 + \epsilon\|\nu_2\|_2^2.$$

So an approximation with exactly k terms is possible, but with cost $1/\epsilon^2$ versus $1/\epsilon$ for the general case.

Almost all sublinear algorithms begin with the observation that if a signal consists of a single heavy hitter, then the trivial encoding of the positions 1 through N with $\log(N)$ bits, referred to as a bit tester, can identify the position of the heavy hitter, as in the following. (This is a specific solution to a much more general problem known as *combinatorial group testing*.)

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 7 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 7 \\ 0 \\ 7 \\ 0 \end{pmatrix}.$$

The second observation in the design and analysis of other sublinear algorithms is that a number of hash or Bernoulli functions drawn at random from a hash family are sufficient to isolate enough of the heavy hitters, which can then be identified by the bit tester. Depending on the type of error metric desired, the hashing matrix is pre-multiplied by random ± 1 vectors (for the ℓ_2 metric) in order to estimate the signal values. In this case, the measurements are referred to as the COUNT SKETCH in the data stream literature [1] and, without the premultiplication, the measurements are referred to as COUNT MEDIAN [5, 6] and give $\ell_1 \leq C\ell_1$ error guarantees. In addition, the sublinear algorithms are typically greedy, iterative algorithms that recover portions of the heavy hitters with each iteration or that recover portions of the ℓ_2 (or ℓ_1) energy of the residual signal (similar to that we saw in the previous section).

We build upon the COUNT SKETCH design but incorporate the following algorithmic innovations to ensure an optimal number of measurements:

- With a random assignment of N signal positions to $O(k)$ subsignals, we need to encode only $O(N/k)$ positions, rather than N as in the previous approaches. Thus we can reduce the domain size which we encode.
- We use a good error-correcting code (rather than the trivial identity code of the bit tester).
- Our algorithm is an iterative algorithm but maintains a *compound* invariant: in our algorithm, the number of undiscovered heavy hitters decreases at each iteration while, simultaneously, the required error tolerance and failure probability become more stringent. Because there are fewer heavy hitters to find at each stage, we can use more measurements to meet more stringent guarantees.

5. Conclusion

Sparse approximation and its related problem of sparse signal recovery both have rich algorithmic and mathematical solutions, as well as a number of still unanswered questions. Many of those questions are connected to practical applications—how to design feasible measurement systems that meet practical considerations. For these problems, both new mathematical and algorithmic tools must be developed in cooperation with the application experts.

Acknowledgements. ACG is partially supported by grants NSF grant CCF-1161233 and NSF CIF 0910765.

References

[1] M. Charikar, K. Chen, and M. Farach-Colton, *Finding frequent items in data streams*, ICALP, 2002.

[2] Raphaël Clifford, Klim Efremenko, Ely Porat, and Amir Rothschild, *k-mismatch with don't cares*, In ESA (2007), 151–162.

[3] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore, *Compressed sensing and best-term approximation*, Journal of the American mathematical society, **22**(1) (2009), 211–231.

[4] G. Cormode and S. Muthukrishnan, *What's hot and what's not: Tracking most frequent items dynamically*, In Proc. ACM Principles of Database Systems (2003), 296–306.

[5] _____, *Improved data stream summaries: The count-min sketch and its applications*, FSTTCS, 2004.

[6] _____, *Combinatorial algorithms for Compressed Sensing*, In Proc. 40th Ann. Conf. Information Sciences and Systems, Princeton, Mar. 2006.

[7] Yaniv Erlich, Kenneth Chang, Assaf Gordon, Roy Ronen, Oron Navon, Michelle Rooks, and Gregory J. Hannon, *Dna sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis*, Genome Research, **19** (2009), 1243–1253.

[8] A. C. Gilbert, M. J. Strauss, J. A. Tropp, and R. Vershynin, *One sketch for all: fast algorithms for compressed sensing*, In ACM STOC (2007), 237–246.

[9] Anna C. Gilbert, Yi Li, Ely Porat, and Martin J. Strauss. *Approximate sparse recovery: Optimizing time and measurements*, SIAM J. Comput. **41**(2) (2012), 436–453.

[10] J.A. Tropp, *Greed is good: algorithmic results for sparse approximation*, Information Theory, IEEE Transactions on, **50**(10) (2004), 2231–2242.

[11] Y. H. Zheng, N. P. Pitsianis, and D. J. Brady, *Nonadaptive group testing based fiber sensor deployment for multiperson tracking*, IEEE Sensors Journal **6**(2) (2006), 490–494.

[12] Y.H. Zheng, D. J. Brady, M. E. Sullivan, and B. D. Guenther, *Fiber-optic localization by geometric space coding with a two-dimensional gray code*, Applied Optics, **44**(20) (2005), 4306–4314.

Department of Mathematics, University of Michigan, 2074 East Hall, 530 S. Church Street, Ann Arbor, MI USA 48109

E-mail: annacg@umich.edu

A mathematical perspective of image denoising

Miguel Colom, Gabriele Facciolo, Marc Lebrun, Nicola Pierazzo,
Martin Rais, Yi-Qing Wang, and Jean-Michel Morel

Abstract. Digital images are matrices of regularly spaced samples, the pixels, each containing a photon count. Each pixel thus contains a random sample of a Poisson variable. Its mean would be the ideal image value at this pixel. It follows that all images are random discrete processes and therefore "noisy". Ever since digital images exist, numerical methods have been proposed to recover the ideal mean from its random observed value. This problem is obviously ill posed and makes sense only if there is an underlying image model. Inventing or learning from data a consistent mathematically image model is the core of the problem. Images being 2D projections of our complex surrounding visual world, this is a challenging problem, which is nevertheless beginning to find simple but mathematically innovative answers. We shall distinguish four classes of denoising principles, relying on functional or stochastic image models. We show that each of these principles can be summarized in a single formula. In addition these principles can be combined efficiently to cope with the full image complexity. This explains their immediate industrial impact. All current cameras and imaging devices rely directly on the simple formulas explained here. In the past ten years the image quality delivered to users has increased fast thanks to this exemplary mathematical modeling.

As an illustration of the universality and simplicity reached by the theory, most image denoising algorithms discussed in this paper can be tested directly on any digital image at *Image Processing On Line*, http://www.ipol.im/. In this web journal, each paper contains a complete algorithmic description, the corresponding source code, and can be run online on arbitrary images.

Mathematics Subject Classification (2010). Primary 62H35; secondary 68U10, 94A08.

Keywords. Image denoising, Fourier transform, Wiener estimate, wavelet threshold, discrete cosine transform, oracle estimate, Bayes formula, neighborhood filters, nonlocal methods, neural networks, blind denoising.

1. Introduction

Most digital images and movies are currently obtained by a matrix of sensors counting photons hitting the surface. We shall denote by $\mathbf{i}$ the indices of the matrix elements also called pixels. The value $\tilde{u}(\mathbf{i})$ observed by a sensor at a pixel $\mathbf{i}$ is a Poisson random variable whose mean $u(\mathbf{i})$ would be the ideal image. The difference between the observed image and the ideal image $\tilde{u}(\mathbf{i}) - u(\mathbf{i}) = n(\mathbf{i})$ is called "noise". By a well known property of Poisson random variables, the standard deviation of the noise $n(\mathbf{i})$ is equal to $\sqrt{u(\mathbf{i})}$. On a motionless scene with constant lighting, $u(\mathbf{i})$ can be approached by simply accumulating photons for a long exposure time, and by taking the temporal average of this photon count. Accumulating photon

Proceedings of the International Congress of Mathematicians, Seoul, 2014

impacts on a sensitive surface is therefore the essence of photography. The first Nicéphore Niépce photograph [16] was obtained after an eight hours exposure: it is very noisy, though! A digitization of it can be seen on the left hand side of Figure 1.1. The image in the middle is an attempt to denoise it. The image on the right is the "estimated noise", namely the difference between the noisy image and its denoised version. How was this done will be explained in section 7.

Figure 1.1. Left: A digitization of the first ever photograph by Nicéphore Niépce "View from the Window at Le Gras" ca. 1826 obtained after an eight hours exposure. Middle: an attempt to denoise it. Right: the "estimated noise", namely the difference between the noisy image and its denoised version.

Augmenting the exposure time of the camera amounts to augmenting the expectation $u(\mathbf{i})$ of the number of photons $\tilde{u}(\mathbf{i})$. The number of photons has mean $u(\mathbf{i})$ and variance $u(\mathbf{i})$. Since this variance measures the amount of noise, this implies that noise increases with the exposure. But the means increases faster than the noise. Indeed, the correctly scaled measurement of the noise is the Signal to Noise Ratio (SNR), which is defined by

$$SNR := \frac{\text{Mean}(u(\mathbf{i}))}{\sqrt{\text{Var}(\tilde{u}(\mathbf{i}))}} = \frac{u(\mathbf{i})}{\sqrt{u(\mathbf{i})}} = \sqrt{u(\mathbf{i})}. \tag{1.1}$$

The SNR increases like the square root of the exposure time. So the more photons we have, the better. The solution for getting a quality image, adopted from the beginning by Nicephore Niépce, was therefore to extend the exposure time as much as possible.

Yet, in a long exposure the photographed scene is exposed to variations due to changes in lighting, camera motion, and incidental motions of parts of the scene. For example in the town view of Figure 1.1, the walls on the right and left are bright because the Sun had moved during the eight hours exposure. Nowadays, digital cameras are much faster and capture fast moving objects. But even with a short exposure time, the photograph still risks motion blur on any animated scene. On the other hand, if the exposure time is too short, the image is noisy. Thus the main limitation to any imaging system is noise, regardless of its resolution.

At a first glance, the denoising problem is anyway hopeless: how to estimate the mean $u(\mathbf{i})$ of a random Poisson variable, given only one sample $\tilde{u}(\mathbf{i})$ of this variable? The best estimate of this mean knowing $\tilde{u}(\mathbf{i})$ is of course this unique sample $\tilde{u}(\mathbf{i})$. A glimpse of a solution comes from image formation theory. An optical image u is band-limited [63] and therefore smooth. Thus, one can restore the band-limited image u from its noisy version $\tilde{u}$, as was proposed in 1966 in [33], by imposing a decay to its Fourier spectrum. This classic Wiener-Fourier method multiplies the Fourier transform by optimal coefficients to attenuate the noise. It results in a convolution of the image with a "low-pass" kernel. As we shall see, this reduces the noise, but blurs the image. This is the functional perspective on the subject.

But the band-limitedness of u also implies that the random observed image values $\tilde{u}(\mathbf{j})$ at neighboring pixels $\mathbf{j}$ of a pixel $\mathbf{i}$ are positively correlated with $\tilde{u}(\mathbf{i})$. Thus, these values can be

taken into account to obtain a better estimate of $u(\mathbf{i})$. These values being nondeterministic, Bayesian approaches are relevant and have been proposed as early as 1972 [60]. This opens the stochastic perspective on the subject.

In short, there are two complementary early approaches to denoising, the Fourier-Wiener method, and the Bayesian estimation. A third hint is also given: the denoising of a given pixel value $\tilde{u}(\mathbf{i})$ must involve the values of neighboring pixels $\tilde{u}(\mathbf{j})$. This leads us to the question: where are the extra image samples that could be used to denoise the single sample $\tilde{u}(\mathbf{i})$? This question will lead us a long way. It turns out that, not only neighboring pixels in the same image can be used, but actually even pixels from other images! The mathematical innovation here comes from a non-local, or fully non-local approach to image processing, under the generic name of *neighborhood filters*, *nonlocal filters*, and even *global filters*, involving a whole set of images to denoise one.

These three main perspectives will permit us to review the main algorithmic principles which have been proposed for noise removal. All of them require a noise model, which in most of our study will be the Gaussian white noise (we will explain in the next section why this is not a limitation). The three rough denoising principles sketched above will be further combined into five algorithm classes, each one relying on a single formula.

- **The Fourier-Wiener transform thresholding principle**, section 2 : uses the regularity of the image (reflected by its sparsity in a well-chosen orthonormal transform). For the associated Fourier-Wiener image filters, the assumption is that the Fourier (or cosine transform, or wavelet transform) of the image decays quickly, and therefore faster than white noise, which is homoscedastic over all frequencies. An extreme view of this denoising principle is called "sparsity". According to this popular assumption used in *compressed sensing* [13], the ideal image has a few "sparse" coefficients in the right basis. If that is true, a simple threshold on the transform coefficients (on the right Hilbert basis) maintains the signal and kills most of the noise;
- **The self-similarity principle** and the patch based methods (section 3): The image is self-similar, and one can therefore use other "neighboring" pixels of the same image with the same expected colour to denoise a given pixel. The *neighborhood filters* propose to average the samples with similar colours, thus performing an artificial photon accumulation. This self-similarity principle is enhanced by deciding on the similarity of two pixels $\mathbf{i}$ and $\mathbf{j}$ by comparing two image patches surrounding them.
- **The Bayesian patch denoising principle,** section 4: The Bayesian principle extends the above considerations by giving them an optimal formulation, under the assumption that the patches similar to a given image patch follow a stochastic model.
- **The global denoising principle,** section 5. In this extension of the Bayesian model, not only image patches from the same image, but also image patches from *other images* can be used for image denoising. In its maximal extension, this principle can use literally all images of the world, thus giving an explicit point density function for the patch stochastic model.
- **Global neural denoising**, section 6 learns directly the denoising algorithm by a supervised learning algorithm, again learnt from a huge patch database.
- **Blind denoising**, section 7 is the ultimate achievement of the theory, as it considers denoising the image with a completely flexible noise model, learnt from the image

itself. This is the principle that must be used for old photographs and for degraded digital photographs, for which the noise model is unknown.

2. Fourier-Wiener transform thresholding

The white noise model. In this section and in the rest of the paper we shall adopt a convenient simplification of the noise model. We defined the noise as the difference between the observed image and the ideal image $\tilde{u}(\mathbf{i}) - u(\mathbf{i}) = n(\mathbf{i})$. For large enough values of $u(\mathbf{i})$ this random variable tends to be Gaussian. Furthermore, the Anscombe scalar transform $f(\tilde{u}(\mathbf{i}))$, where f is a special function proposed by Anscombe [2] transforms this Poisson noise with a variance depending on the signal $u(\mathbf{i})$ into a nearly Gaussian variable with fixed variance. By applying the Anscombe transform to the image its noise becomes *white, homoscedastic and Gaussian*. White means that the random value is independent at each pixel, which is true because the fluctuations of the photon numbers hitting each pixel are independent. Homoscedastic means that all pixels noises have the same variance which we will denote by σ^2. This noise model will simplify the discussion without loss of generality.

Classic transform thresholding algorithms use the observation that images are faithfully described by keeping only their large coefficients in a well-chosen basis. By keeping these large coefficients and setting to zero the small ones, noise should be removed and image geometry kept. By any orthogonal transform, the coefficients of an homoscedastic de-correlated noise remain de-correlated and homoscedastic. Here we refer to the classic Fourier, wavelet or cosine transforms, in their discrete version applied to the image matrix viewed as a vector in a large but finite dimension. Applied to digital images, each one of these transforms is an orthogonal transform in the finite dimensional image space. For the Fourier method this amounts to use the DFT (Discrete Fourier Transform). This Fourier method has been extended in the past thirty years to generalized linear space-frequency transforms such as the windowed cosine transform [70] or the many wavelet transforms [50].

The wavelet, or DCT, or Fourier coefficients of a Gaussian white noise with variance σ^2 remain a Gaussian diagonal vector with variance σ^2. The sparsity model assumes that the most "important" image coefficients are much larger than 3σ. Thus, cancelling the coefficients of the noisy image that are smaller (in absolute value) than, for example, 3σ will remove most of the coefficients that are only due to noise, while keeping the large image coefficients. This *sparsity* of image coefficients in certain bases is an empirical observation, used in most denoising and compression algorithms. For example the established image compression algorithms are based on the DCT (in the JPEG 1992 format) or, like the JPEG 2000 format [3], on biorthogonal wavelet transforms [17]. A bit more formally, let $\mathcal{B} = \{g_i\}_{i=1}^{M}$ be an orthonormal basis of $\mathbb{R}^M$, where M is the number of pixels of the noisy image $\tilde{u}$ (handled here as a vector). Then

$$\tilde{u} = \sum_{i=1}^{M} \langle \tilde{u}, g_i \rangle g_i, \quad \text{with} \quad \langle \tilde{u}, g_i \rangle = \langle u, g_i \rangle + \langle n, g_i \rangle \,, \tag{2.1}$$

where $\tilde{u}$, u and n denote respectively the noisy, ideal and noise images and $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product in $\mathbb{R}^M$. Being independent, the noise values $n(\mathbf{i})$ are uncorrelated. They have by assumption zero mean and variance σ^2. We can deduce that the noise coefficients in the new basis remain uncorrelated, with zero mean and variance σ^2. Indeed, denoting by $\mathbb{E}$

the expectation (with respect to the stochastic noise model) we have $\langle n, g_i \rangle = \sum_{\mathbf{r}=1}^{M} g_i(\mathbf{r}) n(\mathbf{r})$ and therefore

$$\begin{aligned}\mathbb{E}[\langle n, g_i \rangle \langle n, g_j \rangle] &= \sum_{\mathbf{r},\mathbf{s}=1}^{M} g_i(\mathbf{r}) g_j(\mathbf{s}) \mathbb{E}[n(\mathbf{r}) n(\mathbf{s})] \\ &= \langle g_i, g_j \rangle \sigma^2 = \sigma^2 \delta[j-i].\end{aligned}$$

In the Fourier-Wiener method, each noisy transform coefficient $\langle \tilde{u}, g_i \rangle$ is modified independently and then the denoised image is estimated by the inverse transform of the new coefficients. Denoting by $a(i)$ the attenuation factor $a(i)$ for the i-th coefficient, the inverse transform yields the denoised version

$$\mathbf{D}\tilde{u} = \sum_{i=1}^{M} a(i) \langle \tilde{u}, g_i \rangle g_i, \tag{2.2}$$

to be compared with (2.1). **D** is often called a *diagonal operator*. The following result, generally attributed to Norbert Wiener, gives the ideal values for $a(i)$:

Theorem 2.1. *The operator $\boldsymbol{D}_{inf}$ minimizing the mean squared error (MSE) $\boldsymbol{D}_{inf} = arg\min_{\boldsymbol{D}} \mathbb{E}\{\|u - \boldsymbol{D}\tilde{u}\|^2\}$ satisfies*

$$a(i) = \frac{|\langle u, g_i \rangle|^2}{|\langle u, g_i \rangle|^2 + \sigma^2}. \tag{2.3}$$

The previous optimal operator attenuates all noisy coefficients. In the methods assuming a "sparsity" for the ideal image u, one further restricts $a(i)$ to be 0 or 1. Then the diagonal operator becomes a projection operator. In that case, a subset of coefficients is kept, and the rest are set to zero. The projection operator that minimizes the MSE under that constraint is obtained with

$$a(i) = \begin{cases} 1 & \text{if } |\langle u, g_i \rangle|^2 \geq \sigma^2, \\ 0 & \text{otherwise.} \end{cases}$$

A *transform thresholding* algorithm therefore keeps the coefficients with a magnitude larger than the noise, while setting to zero the rest. Note that both above mentioned filters are "ideal", or "oracular" operators. Indeed, they use the coefficients $\langle u, g_i \rangle$ of the original image, which are not known. For this reason, such algorithms are called *oracle filters*. The classical *transform threshold filters* must approximate the oracle coefficients by using the observable noisy coefficients. The real denoising method is therefore called *empirical Wiener filter*, because it approximates the unknown original coefficients $\langle u, g_i \rangle$ by invoking the identity

$$\mathbb{E}|\langle \tilde{u}, g_i \rangle|^2 = |\langle u, g_i \rangle|^2 + \sigma^2$$

to replace the optimal attenuation coefficients $a(i)$ by the empirical attenuation coefficients

$$\alpha(i) = \max\left\{0, \frac{|\langle \tilde{u}, g_i \rangle|^2 - c\sigma^2}{|\langle \tilde{u}, g_i \rangle|^2}\right\} \tag{2.4}$$

where c is a parameter, usually larger than one.

The global Fourier basis is not used for denoising. Indeed, modifying Fourier coefficients by the diagonal operator often causes undue oscillation. To avoid this effect, the orthogonal bases are usually more local, of the wavelet or block DCT type. We give now two examples.

The sliding window DCT. The local adaptive filters were introduced by Yaroslavsky and Eden [70] and Yaroslavsky [72]. The noisy image is analyzed in a moving square block, typically with dimensions 8×8. At each position of the block center, its DCT spectrum is computed and modified by using the empirical coefficients (2.4). Finally, an inverse transform is used to estimate only the signal value in the central pixel of the block.

Wavelet thresholding. Let $\mathcal{B} = \{g_i\}_i$ be a wavelet orthonormal basis [49]. The so-called *hard wavelet thresholding method* [26] is a (nonlinear) projection operator setting to zero all wavelet coefficients smaller than a certain threshold. The performance of the method depends on the ability of the basis to approximate the image U by a small set of large coefficients. There has been a strenuous search for wavelet bases adapted to images [52].

Unfortunately, the brutal cancelation of DCT coefficients near the image edges[1] creates small oscillations by the Gibbs phenomenon. Similarly, the undue cancelation of some of the small wavelet coefficients may also cause the appearance of isolated wavelets in flat image regions. These annoying artifacts are sometimes called *wavelet outliers* [27]. They can be partially avoided with the use of a soft thresholding [25],

$$\alpha(i) = \begin{cases} \frac{\langle \tilde{u}, g_i \rangle - sgn(\langle \tilde{u}, g_i \rangle)\mu}{\langle \tilde{u}, g_i \rangle}, & \text{if } |\langle \tilde{u}, g_i \rangle| \geq \mu, \\ 0 \quad \text{otherwise}, \end{cases}$$

which reduces the Gibbs oscillation near image discontinuities.

Several carefully designed orthogonal bases adapt better to image local geometry and discontinuities than wavelets, particularly the "bandlets" [52] and "curvelets" [65]. This tendency to adapt the transform locally to the image is accentuated with the methods adapting a different basis to each pixel, or selecting a few elements or "atoms" from a huge patch dictionary to linearly decompose the local patch on these atoms. This point of view is developed in *sparse coding methods* and the K-SVD algorithm [1, 29, 47].

2.1. A case study: DCT denoising. We shall illustrate transform thresholding by at least one good detailed example. A basic DCT denoising can be drastically improved by several ingredients illustrated in Figure 2.1. This figure shows how the result improves by successively using a better colour space[2], by aggregating [18] the 64 denoised values obtained for each pixel, which is contained in 64 patches with 8×8 dimensions, by making a statistically more correct aggregation of these estimates, and finally by iterating the method, using the first denoised image as "oracle" for applying the Wiener filter a second time. The method is summarized in Algorithm 1. See [32] for an online implementation.

3. The self-similarity principle and the patch based methods

If m noisy independent pixels with the same expected colour are averaged, the noise (namely the variance of the average of these m values) is divided by m. The first application of this

[1] So are called the strong image discontinuities along apparent contours of visible objects.

[2] A colour image is a set of three images (R, G, B) giving scalar values to three chromatic components, Red, Green, Blue. The linear transform improving the denoising performance is simply $Y_0 = (R + G + B)/3, U_0 = \frac{1}{2}(R - B), V_0 = \frac{1}{4}(R + B) - \frac{1}{2}G$, where Y_0 is the luminance, and U_0 and V_0 contain the colour contrast between green and blue and green and red respectively.

Algorithm 1 DCT denoising algorithm. DCT coefficients lower than 3σ are canceled in the first step and a Wiener filter is applied in the "oracle" second step. In colour this strategy is applied to Y_0 . Its attenuation coefficients are also applied to U_o, V_o.

Input: noisy image $\tilde{u}$, σ noise standard deviation, (optional) prefiltered image $\hat{u}_1$ for "oracle" estimation, $h = 3\,\sigma$: threshold parameter.
Output: output denoised image u.
for each patch $\tilde{P}$ of size 8×8 (if $\hat{u}_1$, patch P_1 in $\hat{u}_1$) **do**
  Compute the DCT transform of $\tilde{P}$ (if $\hat{u}_1$, of P_1).
  if $\hat{u}_1$ **then**
    Modify DCT coefficients of $\tilde{P}$ as $\tilde{P}(i) = \tilde{P}(i)\frac{P_1(i)^2}{P_1(i)^2+\sigma^2}$.
  else
    Cancel coefficients of $\tilde{P}$ with magnitude lower than h.
  end if
  Compute the inverse DCT transform obtaining $\hat{P}$.
  Compute the aggregation weight $w_{\tilde{P}} = 1/\#\{\text{number of non-zero } DCT \text{ coefficients}\}$.
end for
for each pixel **i** **do**
  Aggregation: recover the denoised value at each pixel **i** by averaging all values at **i** of all denoised patches $\hat{Q}$ containing **i**, weighted by $w_{\tilde{Q}}$.
end for

very simple denoising principle is the use of accumulation: when the camera and the scene do not move, the larger the photon count, the larger the signal (mean) to noise (standard deviation) ratio. When we only dispose of a single image, some succedaneous of the above averaging principle must be found to compensate for the limited amount of observed photons. A rather trivial idea is to average the closest pixels to a given pixel. This amounts to convolve the image with a fixed radial positive kernel, for example a Gaussian kernel. This approach works only for pixels inside the homogeneous image regions, but not for those in contrasted image regions. A convolution with a Gaussian may reduce the noise, but it makes the image blurry.

Averaging pixels with similar colours. The sigma-filter [43] or neighborhood filter [71] is an elegant solution to avoid this blur risk. Neighborhood filters average nearby pixels of **i**, but under the condition that they have a colour value similar to that of **i**. These filters denoted by NF for neighborhood filter are defined by

$$NF_{h,\rho}\tilde{u}(\mathbf{i}) = \frac{1}{C(\mathbf{i})} \sum_{\mathbf{j}\in B_\rho(\mathbf{i})} \tilde{u}(\mathbf{j})\, e^{-\frac{|\tilde{u}(\mathbf{i})-\tilde{u}(\mathbf{j})|^2}{h^2}}, \tag{3.1}$$

where $B_\rho(\mathbf{i})$ is a ball of center **i** and radius $\rho > 0$, $h > 0$ is the filtering parameter and $C(\mathbf{i}) = \sum_{\mathbf{j}\in B_\rho(\mathbf{i})} e^{-\frac{|\tilde{u}(\mathbf{j})-\tilde{u}(\mathbf{i})|^2}{h^2}}$ is the normalization factor to make the above an averaging filter. The parameter h expresses the required degree of colour similarity between **i** and **j**. The filter (3.1) is so powerful that it has been been reinvented several times and received several names: *σ-filter* [43], *SUSAN filter* [64] and *Bilateral filter* [66].

3.1. Non-local means. The Non-local means filter extends the concept of a neighborhood filter by implicitly assuming a Markov field structure for the image. Its idea stems from the now famous algorithm to synthesize textures from examples [28]. Its Markovian assumption

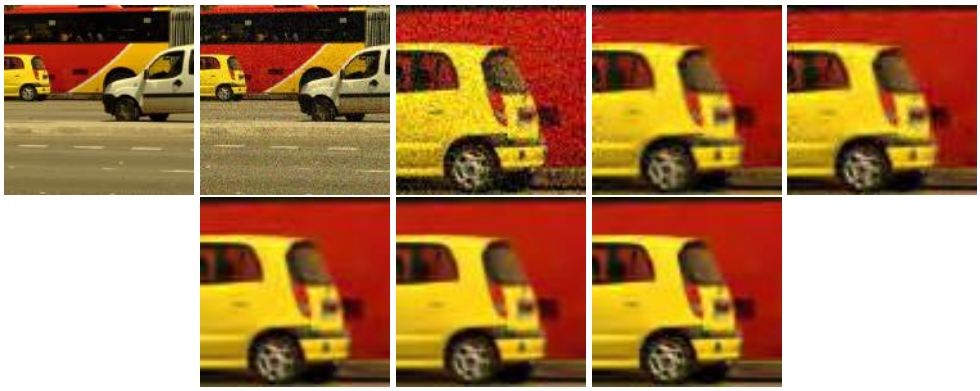


Figure 2.1. Original and noisy images with additive Gaussian white noise; crops of denoised images by Algorithm 1 when incrementally adding the use of a $Y_oU_oV_o$ colour system, uniform aggregation of the 64 estimated values at each pixel, statistically optimal aggregation of the same estimates, and iteration of the Wiener filter with the "oracle" given by the first step. Image quality and SNR increase significantly at each step.

is that, in a textured image, the stochastic model for a given pixel **i** can be predicted from a local image neighborhood P of **i**, which we shall call "patch".

The assumption for recreating new textures from samples is that there are enough pixels **j** similar to **i** in a texture image $\tilde{u}$ to recreate a new but similar texture u. This algorithm goes back to Shannon's theory of communication [63], where it was used for the first time to synthesize a probabilistically correct text from a sample.

An adaptation of the above synthesis principle yields an image denoising algorithm [7][3]. The observed image is the noisy image $\tilde{u}$. The reconstructed image is the denoised image $\hat{u}$. A noisy patch $\tilde{P}$ surrounding a pixel **i** is restored by looking for the patches $\tilde{Q}$ in $\tilde{u}$ with the same dimensions as $\tilde{P}$ and resembling P. Then the restored value $\hat{u}(\mathbf{i})$ is a weighted average of the central values $\tilde{u}(\mathbf{j})$ of the patches resembling P. This defines the"non-local means" algorithm, called "non-local" because it uses patches $\tilde{Q}$ that can be lie far away from $\tilde{P}$, and *even patches taken from other images*.

The underlying self-similarity hypothesis is that for every small patch in a natural image one can find several similar patches in the same image, as illustrated in figure 3.1. Let us now give the formula. NL-means denoises a square reference patch $\tilde{P}$ around **i** of dimension $\kappa \times \kappa$ by replacing it by an average of all similar patches $\tilde{Q}$ in a square neighborhood of **i** of size $\lambda \times \lambda$. To do this, a normalized Euclidean distance between $\tilde{P}$ and $\tilde{Q}$, $d(\tilde{P},\tilde{Q}) = \frac{1}{\kappa^2}\|\tilde{P}-\tilde{Q}\|^2$ is computed for all patches $\tilde{Q}$ is the search neighborhood. Then the weighted average is

$$\hat{P} = \frac{\sum_{\tilde{Q}} \tilde{Q} e^{-\frac{d(\tilde{P},\tilde{Q})^2}{h^2}}}{\sum_{\tilde{Q}} e^{-\frac{d(\tilde{P},\tilde{Q})^2}{h^2}}}. \tag{3.2}$$

The whole method is given in Algorithm 2 and can be tested in IPOL [10].

NL-means works better than the neighborhood filters because the distances of colours between pixels are computed on a patch surrounding the pixel instead of just the central pixel.

[3]See also the related attempts [4, 23, 51, 69].

Algorithm 2 NL-means algorithm.

Input: noisy image $\tilde{u}$, σ noise standard deviation. **Output**: denoised image $\hat{u}$.
Parameters: $\kappa = 3$: patch size, $\lambda = 31$: size of search zone for similar patches, $h = 0.6\,\sigma$: filtering parameter (these values may depend on the noise level)
for each pixel **i** **do**
  Select a square reference patch $\tilde{P}$ around **i** of dimension $\kappa \times \kappa$. Set $\hat{P} = 0$ and $\hat{C} = 0$.
  for each patch $\tilde{Q}$ in a square neighborhood of **i** of size $\lambda \times \lambda$ **do**
    Compute the normalized Euclidean distance $d(\tilde{P}, \tilde{Q}) = \frac{1}{\kappa^2}\|\tilde{P} - \tilde{Q}\|^2$.
    Accumulate $\tilde{Q}e^{-\frac{d(\tilde{P},\tilde{Q})^2}{h^2}}$ to $\hat{P}$ and $e^{-\frac{d(\tilde{P},\tilde{Q})^2}{h^2}}$ to $\hat{C}$.
  end for
  Normalize the average patch $\hat{P}$ by dividing it by the sum of weights $\hat{C}$.
end for
for each pixel x **do**
  Aggregation: recover the denoised value at each pixel **i** by averaging all values at **i** of all denoised patches $\hat{Q}$ containing **i**.
end for

Thus only values of really similar pixels are averaged. This progress is illustrated in Figure 3.2 where the pixel "neighborhoods" have an increasing sophistication: the first result, on an original scanned image, is obtained by a Gaussian convolution. Efficient in flat regions, this filter blurs the edges. The second result is obtained by Yaroslavsky's neighborhood filter: each pixel is replaced by an average of the pixels which are close to it in both the image domain and colour range. The result is much sharper. The last result is obtained by NL-means. The choice of resembling pixels is still more selective. The image differences between original and denoised demonstrate the progress. This difference looks increasingly like noise when the pixel neighborhood becomes more sophisticated. The underlying self-similarity assumption can be formalized by an ergodic assumption, under which NL-means can be proved to converge asymptotically to the noiseless image [4]. The more samples the better, so the algorithm is immediately extendable to video [9]. Figure 3.1 illustrates how NL-means chooses the right weight configuration for each sort of image self-similarity.

4. The Bayesian patch denoising principle

Given u the noiseless ideal image and $\tilde{u}$ the noisy image corrupted with Gaussian noise of standard deviation σ so that

$$\tilde{u} = u + n, \tag{4.1}$$

the conditional distribution $\mathbb{P}(\tilde{u} \mid u)$ is

$$\mathbb{P}(\tilde{u} \mid u) = \frac{1}{(2\pi\sigma^2)^{\frac{M}{2}}} e^{-\frac{||u-\tilde{u}||^2}{2\sigma^2}}, \tag{4.2}$$

where M is the total number of pixels in the image. In order to compute the probability of

[4] It can be proved [7] that if the image is a fairly general stationary and mixing random process, for every pixel **i**, NL-means converges to the conditional expectation of **i** knowing its neighborhood, which is the best Bayesian estimate.

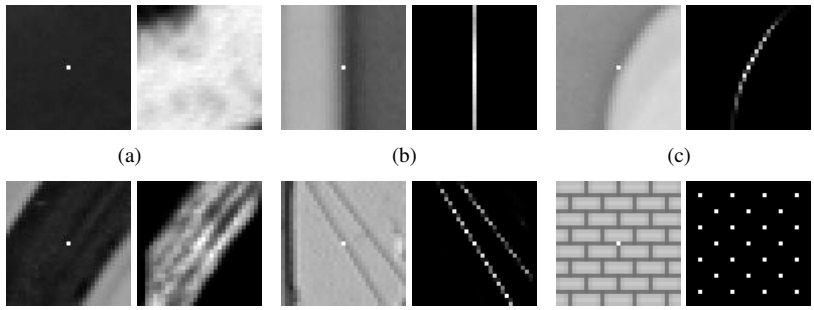


Figure 3.1. On the right-hand side of each pair, one can see the weights in the NL-means average used to estimate a 3×3 patch located in the center of the left image by NL-means.

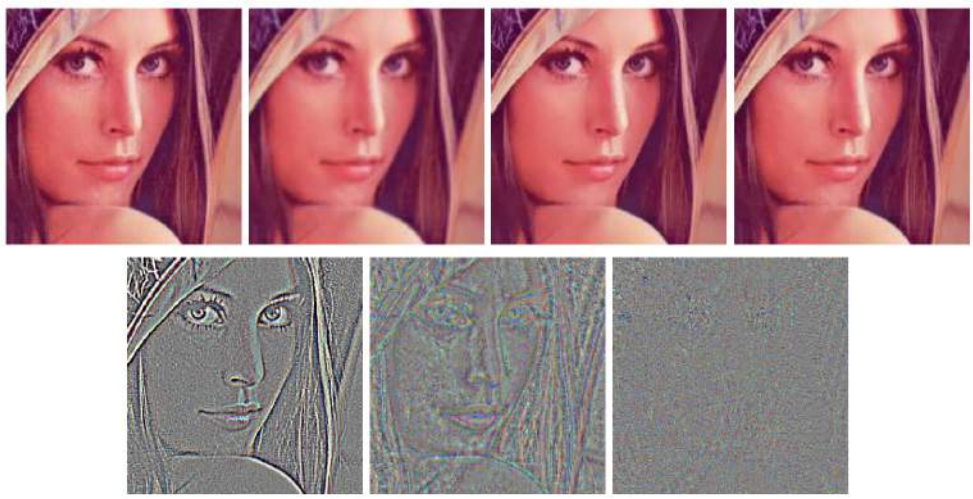


Figure 3.2. A comparison of the efficiency of neighborhood filters. The first row shows a piece of a famous test image (Lena) followed by its denoised version by a Gaussian convolution, a neighborhood filter, and NL-means. The second row shows the difference between the image and its denoised version, which increasingly resembles white noise.

the original image given the degraded one, $\mathbb{P}(u \mid \tilde{u})$, we need a prior on u. In the first models [30], this prior was a parametric Markov random field, specified by its Gibbs distribution. A Gibbs distribution for an image u takes the form

$$\mathbb{P}(u) = \frac{1}{Z} e^{-E(u)/T},$$

where Z and T are constants and E is called the energy function and writes

$$E(u) = \sum_{C \in \mathcal{C}} V_C(u),$$

where $\mathcal{C}$ denotes the set of cliques associated to the image and V_C is a potential function. The maximization of the *a posteriori* distribution writes by Bayes formula

$$\operatorname*{Arg\,max}_u \mathbb{P}(u \mid \tilde{u}) = \operatorname*{Arg\,max}_u \mathbb{P}(\tilde{u} \mid u)\mathbb{P}(u),$$

which is equivalent to the minimization of $-\log \mathbb{P}(u \mid \tilde{u})$,

$$\operatorname*{Arg\,min}_u \|u - \tilde{u}\|^2 + \frac{2\sigma^2}{T} E(u).$$

This energy writes as a sum of local derivatives of pixels in the image, thus being equivalent to a classical Tikhonoff regularization, [30], [6].

Recent Bayesian methods have abandoned as too simplistic the global patch models formulated by a parametric Gibbs energy. Instead, the methods build local non parametric patch models learnt from the image itself, usually as a local Gaussian model around each given patch, or as a Gaussian mixture. The term "patch model" is now preferred to the notion of "clique" previously used for the Markov field methods. But the underlying notion is the same: a "patch" is nothing but a clique. The difference is that the patch model is local and empirical while the clique probability model was usually global and parametric. In the nonparametric local patch models, the patches can become larger, up to an 8×8 size, while the cliques were often confined to very small neighborhoods. Given a noiseless patch P of u with dimension $\kappa \times \kappa$, and $\tilde{P}$ an observed noisy version of P, the same model gives by the independence of noise pixel values

$$\mathbb{P}(\tilde{P}|P) = c \cdot e^{-\frac{\|\tilde{P}-P\|^2}{2\sigma^2}} \tag{4.3}$$

where P and $\tilde{P}$ are considered as vectors with κ^2 components $||P||$ denotes the Euclidean norm of P, and c is the normalizing constant. Knowing $\tilde{P}$, our goal is to deduce P by maximizing $\mathbb{P}(P|\tilde{P})$. Using Bayes' rule, we can compute this last conditional probability as

$$\mathbb{P}(P|\tilde{P}) = \frac{\mathbb{P}(\tilde{P}|P)\mathbb{P}(P)}{\mathbb{P}(\tilde{P})}. \tag{4.4}$$

$\tilde{P}$ being observed, this formula can in principle be used to deduce the patch P maximizing the right term, viewed as a function of P. This is only possible if we know the probability model $\mathbb{P}(P)$. This model will be learnt from the image itself, or from a set of images[5]. For example, once we have obtained (like with NL-means) a group of similar patches Q similar to a given noisy patch P, these patches can be treated as a set of samples of a Gaussian vector. This permits to denoise each observed patch by a Bayesian estimation under this Gaussian model [38]. Let us assume that the patches Q similar to P follow a Gaussian model with (observable, empirical) covariance matrix $\mathbf{C}_P$ and (observable, empirical) mean $\overline{P}$. This means that

$$\mathbb{P}(Q) = c.e^{-\frac{(Q-\overline{P})^t \mathbf{C}_P^{-1}(Q-\overline{P})}{2}} \tag{4.5}$$

From (4.2) and (4.4) we obtain for each observed $\tilde{P}$ the following equivalence of problems:

$$\max_P \mathbb{P}(P|\tilde{P}) \quad \Leftrightarrow \quad \max_P \mathbb{P}(\tilde{P}|P)\mathbb{P}(P)$$

[5]For example [15], [68] or [75] apply a clustering method to the set of patches of a given image before restoration, and [77] applies it to a huge set of patches extracted from many images.

$$\Leftrightarrow \quad \max_P e^{-\frac{\|P-\tilde{P}\|^2}{2\sigma^2}} e^{-\frac{(P-\overline{P})^t \mathbf{C}_{\overline{P}}^{-1}(P-\overline{P})}{2}}$$

$$\Leftrightarrow \quad \min_P \frac{\|P-\tilde{P}\|^2}{\sigma^2} + (P-\overline{P})^t \mathbf{C}_P^{-1}(P-\overline{P}).$$

This expression does not yield an algorithm. Indeed, the noiseless patch P and the patches similar to P are not observable. So we face the same problem as with the oracular Fourier-Wiener filter. Nevertheless, we dispose of the noisy version $\tilde{P}$ and can compute the patches $\tilde{Q}$ similar to $\tilde{P}$. An empirical covariance matrix can therefore be obtained for the patches $\tilde{Q}$ similar to $\tilde{P}$. Furthermore, using (4.1) and the fact that P and the noise n are independent, it is easily checked that

$$\mathbf{C}_{\tilde{P}} = \mathbf{C}_P + \sigma^2 \mathbf{I}; \qquad E\tilde{Q} = \overline{P}. \tag{4.6}$$

If the above empirical estimates are reliable, the maximum *a posteriori* estimation problem finally boils down by (4.6) to the minimization problem:

$$\max_P \mathbb{P}(P|\tilde{P}) \Leftrightarrow \min_P \frac{\|P-\tilde{P}\|^2}{\sigma^2} + (P-\overline{\tilde{P}})^t (\mathbf{C}_{\tilde{P}} - \sigma^2\mathbf{I})^{-1}(P-\overline{\tilde{P}}).$$

Differentiating this quadratic function with respect to P and equating to zero yields the amazingly simple denoising formula

$$\hat{P}_1 = \overline{\tilde{P}} + \left[\mathbf{C}_{\tilde{P}} - \sigma^2\mathbf{I}\right] \mathbf{C}_{\tilde{P}}^{-1}(\tilde{P} - \overline{\tilde{P}}). \tag{4.7}$$

The formula (4.7) gives a direct denoising algorithm, provided we can compute the patch expectations and patch covariance matrices. This is done in [38] by computing empirical means and covariances from the patches similar to a given noisy patch. Since the first such estimate is not accurate, it is natural to iterate the algorithm, so that means and covariances are computed again from denoised patches at the first step. Thus, Algorithm 3 is a self-explanatory application of the single formula (4.7).

As pointed out in [41], the above Nonlocal Bayes algorithm is a Bayesian interpretation (with some generic improvements like the aggregation) of the PCA based algorithm proposed in [76][6].

5. The global patch denoising principle

The most recent denoising methods tend to give up any image model. Indeed, they directly use the observed set of images to denoise a new one. More specifically they denoise image patches by a fully non-local algorithm, in which the patch is compared to a patch model obtained from a large or *very large* patch set, of up to 10^{10} patches. Each patch is denoised by deducing its likeliest estimate from the set of all patches. In the method proposed in [77], this patch space is organized as a Gaussian mixture with about 200 components[7].

[6] See also [24] for a comparison of several local and more global strategies. Non Gaussian, Bayesian models are possible, depending on the patch and noise models. For example [59] treats the case of a local exponential density model for the noisy data.

[7] A similar idea was used in [34] who claim performing a "Scene completion using millions of photographs" to fill in missing parts of a given image.

Algorithm 3 Non local Bayes image denoising

Input: noisy image $\tilde{u}$, σ noise standard deviation. **Output**: denoised image $\hat{u}$.

for all patches $\tilde{P}$ of the noisy image **do**

Find a set $\mathcal{P}(\tilde{P})$ of patches $\tilde{Q}$ similar to $\tilde{P}$.

Compute the expectation $\overline{\tilde{P}}$ and covariance matrix $\mathbf{C}_{\tilde{P}}$ of these patches by

$$\mathbf{C}_{\tilde{P}} \simeq \frac{1}{\#\mathcal{P}(\tilde{P})-1} \sum_{\tilde{Q}\in\mathcal{P}(\tilde{P})} \Big(\tilde{Q}-\overline{\tilde{P}}\Big)\Big(\tilde{Q}-\overline{\tilde{P}}\Big)^t, \qquad \overline{\tilde{P}} \simeq \frac{1}{\#\mathcal{P}(\tilde{P})} \sum_{\tilde{Q}\in\mathcal{P}(\tilde{P})} \tilde{Q}.$$

Obtain the first step estimation $\hat{P}_1 = \overline{\tilde{P}} + \left[\mathbf{C}_{\tilde{P}} - \sigma^2\mathbf{I}\right]\mathbf{C}_{\tilde{P}}^{-1}(\tilde{P}-\overline{\tilde{P}})$.

end for

Obtain the pixel value of the basic estimate image $\hat{u}_1$ as an average of all values of all denoised patches $\hat{Q}_1$ which contain $\mathbf{i}$.

for all patches $\tilde{P}$ of the noisy image **do**

Find a new set $\mathcal{P}_1(\tilde{P})$ of noisy patches $\tilde{Q}$ similar to $\tilde{P}$ by comparing their denoised "oracular" versions Q_1 to P_1.

Compute the new expectation $\overline{\tilde{P}}^1$ and covariance matrix $\mathbf{C}_{\hat{P}_1}$ of these patches:

$$\mathbf{C}_{\hat{P}_1} \simeq \frac{1}{\#\mathcal{P}(\hat{P}_1)-1} \sum_{\hat{Q}_1\in\mathcal{P}(\hat{P}_1)} \Big(\hat{Q}_1-\overline{\tilde{P}}^1\Big)\Big(\hat{Q}_1-\overline{\tilde{P}}^1\Big)^t, \qquad \overline{\tilde{P}}^1 \simeq \frac{1}{\#\mathcal{P}(\hat{P}_1)} \sum_{\hat{Q}_1\in\mathcal{P}(\hat{P}_1)} \tilde{Q}.$$

Obtain the second step patch estimate $\hat{P}_2 = \overline{\tilde{P}}^1 + \mathbf{C}_{\hat{P}_1}\left[\mathbf{C}_{\hat{P}_1} + \sigma^2\mathbf{I}\right]^{-1}(\tilde{P}-\overline{\tilde{P}}^1)$.

end for

Obtain the pixel value of the denoised image $\hat{u}(\mathbf{i})$ as an average of all values of all denoised patches $\hat{Q}_2$ which contain $\mathbf{i}$.

In image denoising, the same idea [45] leads to define the simplest universal method, where a huge set of patches is used to estimate the upper limits a patch-based denoising method will ever reach[8]. The preliminary experiments of this paper involved a set of 20, 000 images [62]. The method, even if certainly not practical, is of exquisite simplicity. Given a clean patch P the noisy patch $\tilde{P}$ with Gaussian noise of standard deviation σ has probability distribution

$$\mathbb{P}(\tilde{P} \mid P) = \frac{1}{(2\pi\sigma^2)^{\frac{\kappa^2}{2}}} e^{-\frac{||P-\tilde{P}||^2}{2\sigma^2}}, \tag{5.1}$$

where κ^2 is the number of pixels in the patch. Then given a noisy patch $\tilde{P}$ its optimal estimator for the Bayesian minimum squared error (MMSE) is by Bayes' formula

$$\hat{P} = \mathbb{E}[P \mid \tilde{P}] = \int \mathbb{P}(P \mid \tilde{P}) P dP = \int \frac{\mathbb{P}(\tilde{P} \mid P)}{\mathbb{P}(\tilde{P})} \mathbb{P}(P) P dP. \tag{5.2}$$

Using a huge set of M natural patches (with a distribution supposedly approximating the real natural patch density), we can approximate the terms in (5.2) by $\mathbb{P}(P)dP \simeq \frac{1}{M}$ and

[8]The results of this paper support the "near optimality of state of the art denoising results", the results obtained by the classic state of the art BM3D algorithm being only 0.1 decibel away from optimality for methods using small patches (typically 8×8.) See also [14].

$\mathbb{P}(\tilde{P}) \simeq \frac{1}{M}\sum_i \mathbb{P}(\tilde{P} \mid P_i)$, which in view of (5.1) yields

$$\hat{P} \simeq \frac{\frac{1}{M}\sum_i \mathbb{P}(\tilde{P} \mid P_i) P_i}{\frac{1}{M}\sum_i \mathbb{P}(\tilde{P} \mid P_i)}. \tag{5.3}$$

Thus the final MMSE estimator is nothing but the exact application of NL-means, denoising each patch by matching it to the huge patch database. The final algorithm is summarized in Algorithm 4. Although this algorithm is optimal, it is not yet fully realizable in our current technology[9].

Algorithm 4 Global Bayesian denoising

Inputs: Noisy image $\tilde{u}$ in vectorial form; very large set of M patches P_i extracted from a large set of noiseless natural images. **Output**: Denoised image $\hat{u}$.
for all patches $\tilde{P}$ extracted from $\tilde{u}$ **do**
Compute the MMSE denoised estimate of $\tilde{P}$

$$\hat{P} \simeq \frac{\sum_{i=1}^{M} \mathbb{P}(\tilde{P} \mid P_i) P_i}{\sum_{i=1}^{M} \mathbb{P}(\tilde{P} \mid P_i)}$$

where $\mathbb{P}(\tilde{P} \mid P_i)$ is known from (5.1).
end for
At each pixel $\mathbf{i}$ get $\hat{u}(\mathbf{i})$ as $\hat{P}(\mathbf{i})$, where the patch P is centered at $\mathbf{i}$.
(optional Aggregation) : for each pixel $\mathbf{j}$ of u, compute the denoised version $\hat{u}_{\mathbf{j}}$ as the average of all values $\hat{P}(\mathbf{j})$ for all patches containing $\mathbf{j}$. (This step in not considered in [45].)

5.1. Comparing visual quality. The visual quality of the restored image is obviously a necessary, if not sufficient, criterion to judge the performance of a denoising algorithm. It permits to control the absence of artifacts and the correct reconstruction of edges, texture and fine structure. Figure 5.1 displays the noisy and denoised images for several classic algorithms for noise standard deviations of 30 (where each colour image is on a scale from 0 to 255). The experiment illustrates that algorithms based on wavelets or DCT, like DCT and BLS-GSM, suffer of a strong Gibbs effect near all image edges. This Gibbs effect is nearly not noticeable in the denoised image by K-SVD which uses a transform method in a learned redundant patch basis, or patch dictionary. The NL-means denoised image has no visual artifacts but is more blurred than those given by BM3D and Non-Local Bayes, that have a clearly superior performance to the rest of the algorithms. The BM3D denoised image has some Gibbs effect near edges, which sometimes degrades the visual quality of the solution. Indeed, the BM3D method is a syncretic method combining the grouping of similar patches with a DCT transform thresholding.

In short, the visual quality of DCT, BLS-GSM and K-SVD is inferior to that of NL-means, BM3D and NL-Bayes, because of strong colour noise low frequencies in flat zones, and of a Gibbs effect.

[9] A clever change of variables in the integral (5.2) found in [53] permits to accelerate the calculation in (5.3) by a 1000 factor, but this is still insufficient!

Figure 5.1. Comparison of visual quality. The noisy image was obtained adding a Gaussian white noise of standard deviation 30. From top to bottom and left to right: original, noisy, DCT sliding window, BLS-GSM, NL-means, K-SVD, BM3D, and Non-local Bayes.

6. Global neural denoising

Though optimal in theory, the global Bayesian denoising formula (5.2) has been recently very well approximated by a neural network learning from an equally huge set of image patches. A feed-forward neural network is a succession of non-linear hidden layers followed by an application-dependent decoder

$$f(\cdot,\theta) = \mathfrak{d} \circ h_n \circ \cdots \circ h_1(\cdot), \quad n \geq 1$$

with

$$\forall 1 \leq l \leq n,\ h_l(z_l) = \mathfrak{a}(W_l z_l + b_l)$$

and

$$\mathfrak{d}(z_n) = W_{n+1} z_{n+1} + b_{n+1}$$

in case of a linear decoder. The parameters θ comprise the connection weights W_l and biases b_l. The activation function $\mathfrak{a}(\cdot)$, typically implemented with the hyperbolic tangent or the logistic function, is applied to its input vector element-wise.

Besides being infinitely differentiable, neural networks can approximate arbitrarily well any continuous function on a compact set [35, 44], thereby making them a candidate for regression tasks

$$\begin{aligned} \theta^* &= \operatorname*{Arg\,min}_{\theta} \mathbb{E}\|f(\tilde{x},\theta) - x\|_2^2 \\ &= \operatorname*{Arg\,min}_{\theta} \mathbb{E}\|f(\tilde{x},\theta) - \mathbb{E}[x|\tilde{x}]\|_2^2 \end{aligned} \tag{6.1}$$

where $(\tilde{x}, x)$ denotes a random pair of observation and its ideal prediction, whose joint behavior is governed by some probability law used to define the expectation in (6.1). Note

	BM3D	NL-Bayes	DNN
$\sigma = 25$	32.53	32.61	**32.88**
$\sigma = 50$	29.20	29.34	**29.72**
$\sigma = 75$	27.28	27.22	**27.95**
$\sigma = 170$	13.84	22.99	**24.56**

Table 6.1. Table comparing two state of the art denoising methods with DNN: the PSNR, qui is a logarithm of the SNR defined in (1.1) measures the image quality (the higher the better).

that although we can sample from it, the underlying probability does not have a closed form in general. Moreover, the function $\theta \mapsto f(\tilde{x}, \theta)$ is not convex, leaving us with little choice but to substitute the expectation with an empirical surrogate and rely on the method of steepest descent [5, 42] to conduct the minimization.

Recently, a set of image denoising neural networks [11] has been shown to outperform BM3D [22] and non-local Bayes [38] at several rather high levels of Gaussian noise for which they were trained. Note that these spin-offs of the original non-local means [7] seek information exclusively inside the noisy image while the neural networks learned to estimate the 17-by-17 patch lying at the center of a noisy 39-by-39 noisy observation by looking at noisy and clean patch pairs gathered from other many images. Table 6.1 is a comparison of these algorithms on a benchmark set and the deep neural networks (DNN) consistently dominate the other two for all the four noise levels.

A look at the output layer of the neural network trained at $\sigma = 25$ (Figure 6.2) reveals a locally oscillating behaviour akin to that of wavelets for those visually meaningful synthesis features. This suggests that a sort of optimal Fourier-Wiener filter is being performed.

This impressive performance is reached with neural networks of four hidden layers, each carrying up to 3000 nodes, thereby requiring a computational cost of several 10^6 operations per pixel. Moreover, their enormous sizes also mean long training time: it could take weeks on a modern GPU platform to train just one neural network [12] under a specific level of noise with tens of millions of example pairs. Although through an investigation of the natural patch distribution, it can be shown [67] that a simple linear transform is readily available to make a single neural network work well across all levels of Gaussian noise, the challenge lying ahead is to scale down such a neural network while preserving its performance.

7. Blind denoising

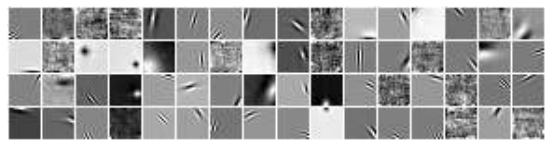
We have shown that all efficient denoising methods boil down to a single formula and to very simple image models. But we assumed a simple noise model, the Gaussian white noise. In this section the focus will be on performing "blind denoising", namely a fully automatic denoising on any digital image.

In most images handled by the public and even by scientists, the noise model is indeed imperfectly known or unknown. Recent progress in noise estimation permits to estimate from a single image a noise model which is simultaneously signal and frequency dependent. We describe here a multiscale denoising algorithm [39] adapted to this broad noise model. This leads to a blind denoising algorithm which can be tested for example on scans of old

(a) Original (b) BM3D

(c) Non local Bayes (d) Neural Network

Figure 6.1. (b), (c) and (d) are the denoised versions of the original image corrupted by some Gaussian noise with standard deviation at 25. The figure shows that a blindly learned algorithm by neural network can outperform all carefully hand-crafted algorithms. Nevertheless the resulting neural network is still unpractical, necessitating tens of millions of connections to denoise a single patch.



(a) output features

Figure 6.2. A random selection of output features. Most of the output features resemble classic wavelets.

photographs, for which the noise model is unknown.

Blind denoising is the conjunction of a noise estimation method followed by the application of an adapted denoising method. Yet, to cope with the broad variety of observed noises, the noise model must be far more comprehensive than the usual white Gaussian noise. Because images have undergone nonlinear operations and filters, a flexible denoising method must cope with a noise model that depends on the signal, but also on the spatial frequency (in technical terms, a coloured noise). The archives of the online executions at the IPOL journal of seven classic denoising methods, namely DCT denoising [72–74], TV denoising [31, 61], K-SVD [40, 48], NL-means [8, 10], BM3D [21, 37], BLS-GSM [58] and NL-Bayes [41] are full with puzzling noisy images.

There are only a few references on blind denoising approaches: Portilla [56], [55], Rabie [57] and Liu, Freeman, Szeliski and Kang [46]. Portilla's method is an adaptation of the BLS-GSM algorithm, which models wavelet patches at each scale by a Gaussian scale mixture (GSM), followed by a Bayesian least square (BLS) estimation for wavelet patches. The "noise clinic" described in this section is based on a noise signal and frequency noise estimator proposed by Colom et al. [19, 20], relying on a Ponomarenko et al. general principle [54] to build a noise patch model.

As evident in its formula (4.7), the NL-Bayes method described in section 4 only requires the knowledge of a local Gaussian patch model and of a Gaussian noise model. We already saw in Algorithm 3 how to estimate the local patch Gaussian model, described by an empirical mean and an empirical covariance. So we only need to hint at how to estimate the covariance matrix of the noise. The noise model being signal dependent, for each intensity $\mathbf{i}$ in the range intensity $[0, 255]$ of the image a noise covariance matrix $\mathbf{C}_{n\mathbf{i}}$ must be estimated. The noise model for each group of patches similar to $\tilde{P}$ will depend on $\tilde{P}$ through their mean $\mathbf{i}$. The reference intensity for the current 3D group $\mathcal{P}(\tilde{P})$ must therefore be estimated to apply (4.7) with the appropriate noise covariance matrix. This intensity is simply estimated as the average of all pixels contained in $\mathcal{P}(\tilde{P})$. So we need to estimate the noise covariance matrices $\{\mathbf{C}_{n\mathbf{i}}\}_{\mathbf{i}\in[0,255]}$. Colom et al., [20], proposed an adaptation of the Ponomarenko et al. [54] method estimating a frequency dependent noise to estimate noise in JPEG images. Given a patch size $\kappa \times \kappa$, the method extracts from the image a set with fixed cardinality of sample blocks with lowest variance, and with mean approximately equal to $\mathbf{i}$. These blocks are therefore likely to contain only noise. They are transformed by a DCT, and an empirical standard deviation of their DCT coefficients is computed. This algorithm computes for every intensity $\mathbf{i}$ with a multi-frequency noise estimate given by a $\kappa^2 \times \kappa^2$ matrix

$$\mathbf{M}_{\mathbf{i}} := \mathbb{E}\left(\mathcal{D}N_{\mathbf{i}}\left(\mathcal{D}N_{\mathbf{i}}\right)^t\right) \tag{7.1}$$

where $\mathcal{D}$ is the $\kappa^2 \times \kappa^2$ matrix of the discrete cosine transform (DCT) and $N_{\mathbf{i}}$ denotes the $\kappa \times \kappa$ stochastic noise patch model at intensity $\mathbf{i}$. This method estimates the variances of the DCT coefficients of noise blocks and not their covariances. The covariance matrices are assumed to be diagonal, since generally the DCT decorrelates the noise.

For a given intensity $\mathbf{i}$, the covariance matrix of the noise is $\mathrm{Cov}(N_{\mathbf{i}}) = \mathbb{E}\left(N_{\mathbf{i}}N_{\mathbf{i}}^t\right)$ which leads to

$$\mathcal{D}\mathrm{Cov}(N_{\mathbf{i}})\mathcal{D}^t = \mathcal{D}\mathbb{E}\left(N_{\mathbf{i}}N_{\mathbf{i}}^t\right)\mathcal{D}^t = \mathbb{E}\left(\mathcal{D}N_{\mathbf{i}}\left(\mathcal{D}N_{\mathbf{i}}\right)^t\right) = \mathbf{M}_{\mathbf{i}} \tag{7.2}$$

thanks to (7.1). The DCT being an orthogonal transform , from (7.2) we get $\mathrm{Cov}(N_{\mathbf{i}}) = \mathcal{D}^t\mathbf{M}_{\mathbf{i}}\mathcal{D}$.

We shall apply the blind denoising to a real noisy image for which no noise model was available. To illustrate the algorithm structure and its action, we present the noisy input image, the denoised image, the difference image = noisy - denoised, the average noise curve over high frequencies, and the average noise curve over low frequencies. The results are shown in Figure 7.1. As the noise curves illustrate, the noise is frequency and signal dependent.

Results on old photographs. Scanned old photographs form a vast image corpus for which the noise model can't be anticipated. The noise is chemical, generally with big grain and further altered by the scanning and JPEG encoding. Figure 7.2 shows results obtained by the Noise Clinic over this kind of noisy images. The results compare well with those obtained with *blind BLS-GSM* [55, 56], another state-of-the-art blind denoising algorithm.

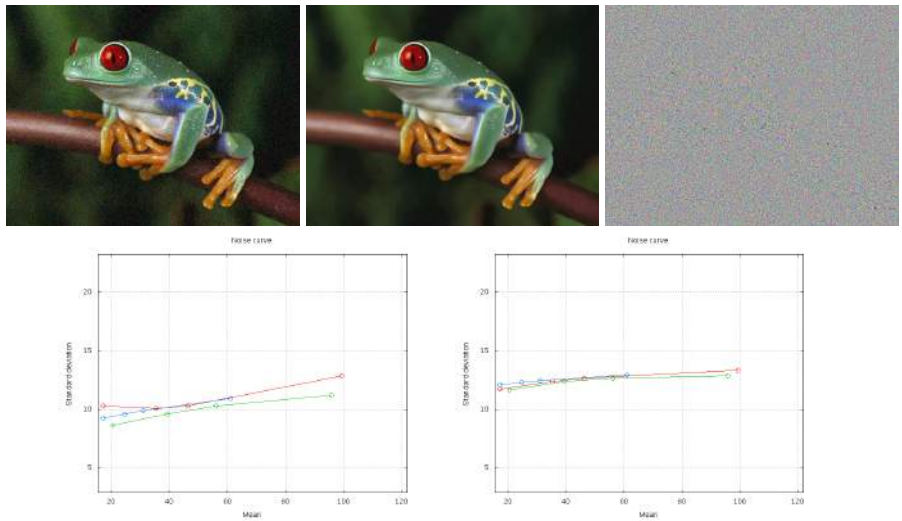


Figure 7.1. Top: Illustration of blind denoising of a JPEG image, the "Frog" image. It is advised to zoom in the high quality pdf file to see detail. Left, noisy image, middle denoised image and right, difference image. Bottom: noise variance estimation of the "Frog" image, as a function of the image value and of the local spatial DCT frequency . Left: average of the low frequency curves in the DCT. Right: average of high frequencies.

Figure 7.2. Blind denoising results on two old photographs. The first is a portrait of young Marylin Monroe. The second is a detail of a group photograph at the Solvay conference, 1927. For both, a crop of a scan of the original image is followed result of the Noise Clinic. It is advised to zoom in the pdf to see image details.

8. Conclusion

Fifty years effort have ended up with denoising methods that can be fully described with six short formulas that guarantee optimality for a definite image model: these formulas are : (2.2) and (2.4) for the Wiener-Fourier transform thresholding assuming an image sparsity model; (3.1) for the neighborhood filter and (3.2) for NL-means, both assuming a self-similarity model; (4.7) for nonlocal Bayes, which assumes again an image self-similarity and local Gaussian behavior for patches. Finally the single formula (5.3) for global Bayesian denoising, which is asymptotically optimal given a (virtually infinite) sample set of image patches. The

global (Bayesian or neural) methods bypass the question of a mathematical image model by using an *in extenso* model, namely *all* image patches of the world. For this precise reason, they are still impractical. On the other hand, the best simple image models obtain a denoising performance equivalent to global methods. This is encouraging for mathematical modeling! But are the three main image mathematical models compatible? The answer is yes: the Bayesian self-similarity image model (Nonlocal Bayes) combines the three main principles. Indeed, the Bayesian local estimate of a patch is a diagonal operator on the patch basis given by the local Gaussian model. Similarly, a recent method, *dual domain denoising* [36], also shows excellent performance by alternating and iterating a neighborhood filter with a DCT transform thresholding.

Acknowledgements. Research partially financed by the Office of Naval research under grant N00014-97-1-0839, DxO-Labs, Centre National d'Etudes Spatiales (CNES, MISS project), the European Research Council, advanced grant "Twelve labours", and the Spanish Ministerio de Ciencia e Innovación under grant TIN2011-27539.

References

[1] M. Aharon, Michael Elad, and A. Bruckstein, *K-SVD: Design of dictionaries for sparse representation*, IEEE Transactions on Image Processing (2005), 9–12.

[2] F. J. Anscombe. *The transformation of Poisson, binomial and negative-binomial data* , Biometrika, **35**(3) (1948), 246–254.

[3] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, *Image coding using wavelet transform*, IEEE Transactions on Image Processing **1**(2) (1992), 205–220.

[4] S.P. Awate and R.T. Whitaker, *Unsupervised, information-theoretic, adaptive image filtering for image restoration*, IEEE Trans. PAMI **28**(3) (2006), 364–376.

[5] L. Bottou. *Large-scale machine learning with stochastic gradient descent*, In Proc. Int. Conf. Computational Statistics, pp. 177–186, Springer, 2010.

[6] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31, Springer Verlag, 1999.

[7] A. Buades, B. Coll, and J. M. Morel, *A review of image denoising algorithms, with a new one*, Multiscale Modeling & Simulation **4**(2) (2005), 490–530.

[8] _____, *A non local algorithm for image denoising*, IEEE Computer Vision and Pattern Recognition **2** (2005), 60–65.

[9] _____, *Nonlocal image and movie denoising*, International Journal of Computer Vision **76**(2) (2008), 123–139.

[10] _____, *Non-local means denoising*, Image Processing On Line **1** (2011).

[11] H. Burger, C. Schuler, and S. Harmeling, *Image denoising: Can plain neural networks compete with BM3D?*, In IEEE Conf. Computer Vision and Pattern Recognition, pp. 2392–2399, 2012.

[12] H.C. Burger, *Modelling and Learning Approaches to Image Denoising*. PhD thesis, Eberhard Karls Universität Tübingen, Wilhelmstr. 32, 72074 Tübingen, 2013.

[13] E.J. Candès and M.B. Wakin, *An introduction to compressive sampling*, Signal Processing Magazine, IEEE **25**(2) (2008), 21–30.

[14] P. Chatterjee and P. Milanfar, *Is denoising dead?*, IEEE Transactions on Image Processing **19**(4) (2010), 895–911.

[15] _____, *Patch-based near-optimal image denoising*, IEEE Transactions on Image Processing, 2011.

[16] C. Chevalier, G. Roman, and J.N. Niepce, *Guide du photographe*, C. Chevalier, 1854.

[17] A. Cohen, I. Daubechies, and J.C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Communications on pure and applied mathematics **45**(5) (1992), 485–560.

[18] R.R. Coifman and D.L. Donoho, *Translation-invariant de-noising*, Lecture Notes In Statistics, pp. 125–125, 1995.

[19] A. Colom, M. Buades, *Analysis and extension of the Ponomarenko et al. method, estimating a noise curve from a single image*, Image Processing On Line **3** (2013), 173–197.

[20] M. Colom, M. Lebrun, A. Buades, and J.M. Morel, *A non-parametric approach for the estimation of intensity-frequency dependent noise*, submitted, 2014.

[21] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, *Image denoising by sparse 3D transform-domain collaborative filtering*, IEEE Transactions on Image Processing **16**(8) (2007).

[22] _____, *Image restoration by sparse 3D transform-domain collaborative filtering* , In Electronic Imaging, 2008.

[23] J.S. De Bonet, *Noise reduction through detection of signal redundancy*, Rethinking artificial intelligence, 1997.

[24] J. Dalalyan A. Deledalle, C.A. Salmon, *Image denoising with patch based PCA: local versus global*, In Proceedings of the British Machine Vision Conference, pp. 25.1–25.10, 2011.

[25] D.L. Donoho, *De-noising by soft-thresholding*, IEEE Transactions on Information Theory **41**(3) (1995), 613–627.

[26] D.L. Donoho and J.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81**(3) (1994), 425–455.

[27] S. Durand and M. Nikolova, *Restoration of wavelet coefficients by minimizing a specially designed objective function*, In Proc. IEEE Workshop on Variational, Geometric and Level Set Methods in Computer Vision, pp. 145–152, 2003.

[28] A.A. Efros and T.K. Leung, *Texture synthesis by non-parametric sampling*. In International Conference on Computer Vision, volume 2, pp. 1033–1038, 1999.

[29] M. Elad and M. Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Transactions on Image Processing **15**(12) (2006), 3736–3745.

[30] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*, IEEE Transactions on Pattern Analysis and Machine Intelligence **6** (1984), 721–741.

[31] P. Getreuer, *Rudin-Osher-Fatemi total variation denoising using split Bregman*, Image Processing On Line **2** (2012).

[32] Guillermo Sapiro Guoshen Yu, *DCT image denoising: a simple and effective image denoising algorithm*, Image Processing On Line **1** (2011).

[33] S.R.J.L. Harris, *Image evaluation and restoration*, Journal of the Optical Society of America **56**(5) (1966), 569–570.

[34] J. Hays and A.A. Efros, *Scene completion using millions of photographs*, In ACM Transactions on Graphics, volume 26, p. 4, 2007.

[35] K. Hornik, M. Stinchcombe, and H. White. *Multilayer feedforward networks are universal approximators*, Neural Networks **2**(5) (1989), 359–366.

[36] C. Knaus and M. Zwicker, *Dual-domain image denoising*, In IEEE International Conference on Image Processing, pp. 440–444, 2013.

[37] M. Lebrun, *An analysis and implementation of the BM3D image denoising method*, Image Processing On Line **2** (2012).

[38] M. Lebrun, A. Buades, and J. M. Morel, *A nonlocal Bayesian image denoising algorithm*, SIAM Journal on Imaging Sciences **6**(3) (2013), 1665–1688.

[39] M. Lebrun, M. Colom, and J.M. Morel, *The noise clinic: a universal blind denoising algorithm*, submitted, 2014.

[40] M. Lebrun and A. Leclaire, *An implementation and detailed analysis of the K-SVD image denoising algorithm*, Image Processing On Line **2** (2012).

[41] Marc Lebrun, Antoni Buades, and Jean-Michel Morel, *Implementation of the "Non-Local Bayes" (NL-Bayes) image denoising algorithm*, Image Processing On Line, **3** (213), 1–42. http://dx.doi.org/10.5201/ipol.2013.16.

[42] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, *Efficient backprop*, In Neural networks: Tricks of the trade, pp. 9–50. Springer, 1998.

[43] J.S. Lee, *Digital image smoothing and the sigma filter*, Computer Vision, Graphics, and Image Processing **24**(2) (1983), 255–269.

[44] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural Networks **6**(6) (1993), 861–867.

[45] A. Levin and B. Nadler, *Natural image denoising: Optimality and inherent bounds*, In IEEE Conference on Computer Vision and Pattern Recognition (2011), 2833–2840.

[46] Liu, W. Freeman, R. Szeliski, and S. Kang, *Automatic estimation and removal of noise from a single image*, IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2) (February 2008), 299–314.

[47] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, *Non-local sparse models for image restoration*, In International Conference on Computer Vision, (2009), 2272–2279.

[48] J. Mairal, G. Sapiro, and M. Elad, *Learning multiscale sparse representations for image and video restoration*, SIAM Multiscale Modeling and Simulation **7**(1) (2008), 214–241.

[49] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic press, 1999.

[50] Y. Meyer, *Wavelets-algorithms and applications*, Wavelets-Algorithms and applications Society for Industrial and Applied Mathematics Translation., p. 142, 1, 1993.

[51] E. Ordentlich, G. Seroussi, S. Verdu, M. Weinberger, and T. Weissman, *A discrete universal denoiser and its application to binary images*, In International Conference on Image Processing, volume 1, 2003.

[52] E. Le Pennec and S. Mallat, *Geometrical image compression with bandelets*, In Proceedings of the SPIE 2003, volume 5150, pp. 1273–1286.

[53] N. Pierazzo and M. Rais, *Boosting shotgun denoising by patch normalization*, In IEEE International Conference on Image Processing (2013).

[54] N. Ponomarenko, V. Lukin, K. Egiazarian, and J. Astola, *A method for blind estimation of spatially correlated noise characteristics*, In IS&T/SPIE Electronic Imaging, pp. 753208–753208, International Society for Optics and Photonics, 2010.

[55] J. Portilla, *Blind non-white noise removal in images using Gaussian scale mixtures in the wavelet domain*, Benelux Signal Processing Symposium, 2004.

[56] ———, *Full blind denoising through noise covariance estimation using Gaussian scale mixtures in the wavelet domain*, IEEE International Conference on Image Processing **2** (2004), 1217–1220.

[57] T. Rabie, *Robust estimation approach for blind denoising*, IEEE Transactions on Image Processing **14**(11) (November 2005), 1755–1765.

[58] Boshra Rajaei, *An Analysis and Improvement of the BLS-GSM Denoising Method*, Image Processing On Line **4** (2014), 44–70.

[59] M. Raphan and E.P. Simoncelli, *An empirical Bayesian interpretation and generalization of NL-means*, Technical report, TR2010-934, Computer Science Technical Report, Courant Inst. of Mathematical Sciences, New York University, 2010.

[60] W.H. Richardson, *Bayesian-based iterative method of image restoration*, Journal of the Optical Society of America **62**(1) (1972), 55–59.

[61] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D **60** (1992), 259–268.

[62] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, *Labelme: a database and web-based tool for image annotation*, International journal of computer vision **77**(1) (2008), 157–173.

[63] C.E. Shannon, *A mathematical theory of communication*, ACM SIGMOBILE Mobile Computing and Communications Review **5**(1) (2001), 3–55.

[64] S.M. Smith and J.M. Brady, *SUSANA new approach to low level image processing*, International Journal of Computer Vision **23**(1) (1997), 45–78.

[65] J.L. Starck, E.J. Candes, and D.L. Donoho, *The curvelet transform for image denoising*, IEEE Transactions on Image Processing **11**(6) (2002), 670–684.

[66] C. Tomasi and R. Manduchi, *Bilateral filtering for gray and color images*, In Computer Vision, 1998. Sixth International Conference on, pp. 839–846, 1998.

[67] Y. Q. Wang and J. M. Morel, *Can a single image denoising neural network handle all levels of Gaussian noise?*, IEEE Signal Processing Letters, 2014. to appear.

[68] ———, *SURE guided Gaussian mixture image denoising*, SIAM Journal on Imaging Sciences **6**(2) (2013), 999–1034.

[69] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and MJ Weinberger, *Universal discrete denoising: Known channel*, IEEE Transactions on Information Theory **51**(1) (2005), 5–28.

[70] L. Yaroslavsky and M. Eden, Fundamentals of Digital Optics, 2003.

[71] L. P. Yaroslavsky, *Digital Picture Processing*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1985.

[72] ———, *Local adaptive image restoration and enhancement with the use of DFT and DCT in a running window*, In Proceedings of SPIE, volume 2825, pp. 2–13, 1996.

[73] L.P. Yaroslavsky, K.O. Egiazarian, and J.T. Astola, *Transform domain image restoration methods: review, comparison, and interpretation*, In Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, volume 4304, pp. 155–169, May 2001.

[74] G. Yu and G. Sapiro, *DCT image denoising: a simple and effective image denoising algorithm*, Image Processing On Line, 1, 2011.

[75] G. Yu, G. Sapiro, and S. Mallat, *Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity*, IEEE Transactions on Image Processing **21**(5) (2012), 2481–2499.

[76] L. Zhang, W. Dong, D. Zhang, and G. Shi, *Two-stage image denoising by principal component analysis with local pixel grouping*, Pattern Recognition **43**(4) (2010), 1531–1549.

[77] D. Zoran and Y. Weiss, *From learning models of natural image patches to whole image restoration*, International Conference on Computer Vision, 2011.

CMLA, ENS Cachan, 61 av. du Psdt Wilson, 94235 Cachan Cedex France
E-mail: morel@cmla.ens-cachan.fr

Scaling in kinetic mean-field models for coarsening phenomena

Barbara Niethammer

Abstract. We consider two paradigms of coarsening systems in materials science, Ostwald Ripening and Grain Growth. Experimental observations suggest that for large times such systems evolve in a universal statistically self-similar fashion. One approach to capture this behaviour is to utilize kinetic mean-field models for the particle size distributions. We review recent progress in the derivation and the analysis of such equations for our two model examples.

Mathematics Subject Classification (2010). 82C26, 35C06.

Keywords. Coarsening, Ostwald Ripening, Grain growth, scaling hypothesis.

1. Introduction

Two fundamental examples of coarsening phenomena in materials science are Ostwald Ripening and Grain Growth. Ostwald Ripening occurs as the last stage of numerous phase transition processes that appear when, due to a change in temperature or pressure for example, the energy of a multicomponent system prefers two different phases of a material such that a homogeneous mixture becomes unstable and separates into two stables phases. Typical examples are phase separation in binary alloys upon cooling or the formation of liquid droplets in a supersaturated vapor.

If one of the phases has smaller volume fraction, this minority phase appears in the form of small droplets, that first grow from a uniform background supersaturation. In the late stage of the phase transition when the supersaturation has become small, surface energy kicks in and the droplets start to interact in order to minimize their total surface area. In so-called diffusion controlled Ostwald Ripening, the limiting mechanism of this interaction is mass transfer by diffusion. Mass is transferred from the smaller particles, that have relatively large surface area, to the larger ones which have relatively small surface area. As a consequence larger particles grow, smaller ones shrink and disappear. This coarsening process is commonly known as Ostwald Ripening, named after Wilhelm Ostwald who was the first to describe and explain this phenomenon.

Grain growth on the other hand denotes the coarsening of grains in a polycrystalline material. Roughly speaking, such a polycrystal appears if a metal melt is cooled below the melting temperature. First, small nuclei of solid crystals are created with a random orientation of their crystal lattice. These nuclei grow, touch each other and finally build a solid block composed of crystals in different orientations. Due to this mismatch of grain orientations

Proceedings of the International Congress of Mathematicians, Seoul, 2014

the surfaces where crystals touch carry surface energy. The late stage in the evolution of the polycrystal is driven by the reduction of this surface energy and is facilitated by migration of atoms along the grain boundaries. Also here, one observes that smaller grains shrink and disappear, while larger ones grow.

Both, Ostwald Ripening and grain growth serve as paradigms for the so-called scaling hypothesis. Based on experimental observations one expects that after some transient stage, the system of particles, i.e. droplets or grains, forgets about its initial configuration and evolves in a universal statistically self-similar fashion. The latter means that the average particle size follows a simple power law in time and the size distribution of droplets rescaled with respect to this mean size develops into a unique stationary form.

In the following we will discuss how one can investigate this self-similar long-time behaviour within a kinetic mean-field type model. The strategy is, similar as in gas dynamics, to derive from a well-established 'microscopic model', that is a continuum model for the evolution of the individual particles, an evolution equation for the particle size distribution and analyze the long-time behaviour of solutions. In analogy to the Boltzmann equation such models are often loosely called kinetic models. The difficulties one faces are at least two-fold. To derive a closed model for the size distribution one typically needs to consider a certain small parameter limit in which interactions are weak, or one needs to make some closure assumptions that are in general hard to justify. The analysis of the long-time behaviour of solutions of the resulting evolution equation poses additional mathematical challenges, since the inherent difficulty is that the equations are typically non-local such that well-established mathematical tools such as comparison principles cannot be applied.

In this article we review progress that has been obtained in recent years for the two phenomena described above. While the theory for Ostwald Ripening started long ago with the classical work of Lifshitz, Slyozov and Wagner [20, 38] and is by now well-developed, grain growth is much less understood. We discuss here some attempts to investigate two-dimensional grain growth via kinetic models. We will also briefly connect the results to a related problem, that is self-similar long-time behaviour in Smoluchowski's coagulation equation.

2. Ostwald Ripening

As indicated above, one can in general expect to justify a kinetic mean-field model in cases where one has in a certain sense weak interaction between particles. Ostwald Ripening is accessible to a mathematical analysis in a small parameter regime that is still relevant for applications which is the regime where the volume fraction of particles is small. In this case one observes in experiments that particles are approximately shaped as balls and essentially do not move in space. Thus we can characterize a particle by a single quantity, its radius, and to investigate statistical self-similarity we can study the evolution of the particle radius distribution. In the following we will always assume that the volume fraction of the particles is small; this regime is also called the dilute regime.

The Mullins-Sekerka evolution. Our starting point for the analysis of Ostwald Ripening is a free boundary problem that is known in the metallurgical literature as the Mullins-Sekerka model. We consider a simplified version of this model that preserves the spherical shape of particles which as described before is appropriate in the dilute regime. In this model particles,

called P_i, are distributed in a domain $\Omega \subset \mathbb{R}^3$ and are characterized by their fixed centers $X_i \in \Omega$ and their radii $R_i(t)$. Particles interact by diffusion, but in late-stage coarsening we can assume that mass exchange between particles is much faster than the growth of the interfaces. Hence we can use a quasi-steady approach, that is we assume that the potential u relaxes at each time instantaneously to equilibrium. This gives that for each time t the potential $u = u(x,t)$ solves

$$\begin{aligned} \Delta u &= 0 && \text{in } \Omega \setminus \cup_i \overline{P_i} \\ u &= \frac{1}{R_i} && \text{on } \partial P_i \,. \end{aligned} \tag{2.1}$$

The second equation in (2.1) is the well-known Gibbs-Thomson law which accounts for surface tension. To define the potential uniquely, we couple (2.1) with a no-flux condition on the boundary, that is

$$\frac{\partial u}{\partial \vec{n}} = 0 \qquad \text{on } \partial\Omega \,. \tag{2.2}$$

We easily convince ourselves that if all particles have the same size, the potential u is constant (indeed equal to the inverse radius of the particles). However, if particles have different sizes, this induces gradients in the potential and these gradients drive the system towards a state of lower energy. The Gibbs-Thomson law in (2.1) implies that u is large at small particles which have large surface area compared to their volume, and small at large particles. Hence, mass diffuses from the small to the large particles. The growth rate of a particle is simply given by the total flux towards the particle, that is

$$\frac{d}{dt}\Big(\frac{4\pi R_i^3}{3}\Big) = \int_{\partial P_i} \frac{\partial\, u}{\partial \vec{n}}\, dS \,, \tag{2.3}$$

where here $\vec{n}$ denotes the outer normal to the particle.

It is not difficult to show that if we start with a finite number of particles, which do not overlap, the problem (2.1)-(2.2) is well-posed and depends Lipschitz-continuously on the radii of the particles. As a consequence, the full time-dependent system (2.1)-(2.3) is well-posed for short times. We can extend such a local solution up to a time when a particle vanishes or when two particles touch. In the first case we just eliminate the particle and continue with the remaining ones. In this way we obtain a continuous in time, piecewise smooth solution. In the second case, where particles touch, there is no way to extend the solution in a reasonable way. In fact, the simplifying assumption that particles are spherical is not a good approximation when particles are close.

However, we are interested in the dynamics of a large set of particles with small volume fraction, and we expect that the event that particles touch is rare if it occurs at all. Hence it is plausible that it does not have an influence on the global behavior of the system. As we shall see, the latter is true to leading order, but not if one is interested in higher order effects. We will return to this issue later in this section.

As long as the evolution is well-posed we easily verify that it preserves the total volume of the particles and decreases the surface energy. Indeed, we have

$$\frac{d}{dt} \sum_i R_i^3 = 0 \tag{2.4}$$

and

$$\frac{d}{dt}\sum_i R_i^2 = -\frac{\pi}{2}\int_\Omega |\nabla u|^2\,dx \le 0\,. \tag{2.5}$$

In contrast to other curvature driven evolutions, such as the mean curvature flow, the Mullins-Sekerka evolution (2.1)-(2.3) is nonlocal. More precisely, the evolution of the radius of one particle depends on all the other particles in the system, since all particles interact via the potential u. A key difficulty is that a priori the interaction range between particles is large due to the slow decay of the fundamental solution of Laplace's equation. The challenge is to derive the effective growth law of a particle in a sea of surrounding particles.

The leading order mean-field theory (LSW-theory). The classical LSW-theory of Ostwald Ripening is based on the idea that in the dilute regime the particle size is much smaller than the typical distance between the nearest neighbours. Hence one can assume that one particle interacts with all the others only through a common spatially constant mean-field $u_\infty(t)$. We then solve for particle P_i

$$\begin{aligned} -\Delta u &= 0 && \text{in } \mathbb{R}^3\backslash\overline{P_i} \\ u &= \frac{1}{R_i} && \text{on } \partial P_i \\ u &\to u_\infty && \text{as } |x|\to\infty\,, \end{aligned} \tag{2.6}$$

whose solution is given by

$$u(x,t) = u_\infty + \frac{1 - R_i u_\infty}{|x - X_i|}\,.$$

Using this solution in (2.3) we obtain the simple law

$$\frac{d}{dt}\Big(\frac{4\pi}{3}R_i^3\Big) = R_i u_\infty - 1\,, \tag{2.7}$$

that is a particle grows if its radius is larger than $\frac{1}{u_\infty(t)}$, the critical radius, while it shrinks if it is smaller. So far, we have not specified u_∞. In the above approximation we have not yet taken into account that the evolution preserves the total volume of the particles. This constraint determines u_∞ and implies that

$$u_\infty = \frac{\sum_{i\,:R_i>0} 1}{\sum_i R_i} = \frac{1}{\text{mean radius}}\,, \tag{2.8}$$

that is in this approximation the critical radius is just the mean radius. Recall that in the coarsening picture the critical radius typically increases, so that over time more and more particles start to shrink and finally disappear.

Based on (2.7) we can now derive an equation for the one-particle number density, that is the expected number of particles with radius R in $(R, R+dR)$, which we denote by $f = f(R,t)$. The system (2.7)-(2.8) translates without further approximation into the following evolution law for f:

$$\partial_t f + \partial_R\Big(\frac{1}{R^2}(Ru_\infty(t) - 1)\,f\Big) = 0 \tag{2.9}$$

with

$$u_\infty(t) = \frac{\int_0^\infty f(R,t)\,dR}{\int_0^\infty Rf(R,t)\,dR} =: \frac{1}{\langle R\rangle}\,. \tag{2.10}$$

On the level of equations (2.9)-(2.10) we can now investigate statistical self-similarity. More precisely, we can ask whether the equation has specific self-similar solutions, and if so, given some initial data f_0, whether solutions with such data converge towards a self-similar profile or not.

One can easily check that the equation has a scale invariance $R \sim t^{1/3}$ which is inherited from the Mullins-Sekerka evolution. Furthermore it is also not difficult to find that (2.9) has self-similar solutions, but not only one but a one-parameter family of the form $f(R,t) = t^{-4/3}F_a(R/t^{1/3})$ with $u_\infty = (at)^{1/3}$ and $a \in (0, \frac{4}{9}]$. All of the self-similar profiles F_a have compact support, $F_{4/9}$ is smooth, the other ones behave like power laws at the end of their support. Interestingly, while all these solutions were found in the work by Lifshitz and Slyozov [20], Wagner [38] found only the smooth one and ignores the others, seemingly due to a computational error. As a consequence of this, Wagner concludes that this solution that he thought to be unique characterizes the large-time behaviour of all solutions. Lifshitz and Slyozov on the other hand, argued that only the smooth self-similar solution is stable. To do this, however, they took effects into account that have been neglected in equation (2.9), namely the collision, or 'encounters' as they call it, of particles. As a consequence of these predictions they obtain universal growth rates of the coarsening process, such as for example that the mean radius evolves as $\left(\frac{4}{9}t\right)^{1/3}$.

Shortcomings of the LSW-theory. After the LSW theory was published many experiments were undertaken to test the validity of its predictions. However, it turned out that the agreement with experiments was not very good. Typically one observes much larger coarsening rate and much broader size distributions (see for example the excellent reviews [36, 37]). In addition to these discussions, also the predictions of universal self-similar behaviour of solutions within the LSW model were the subject of a vigorous discussion in the applied literature [2–4, 17, 18].

It took some time until all of these issues were investigated via a more rigorous mathematical analysis, but interestingly enough, different groups came up with the same conclusion around the same time. More precisely, it was predicted via asymptotic analysis in [13] and in a mathematically rigorous way in [30] (see also [5] for a related model) that the long-time behaviour of solutions to the LSW model is not universal, but depends on the contrary sensitively on the initial data, more precisely on the behaviour at the end of the support or in other words on the largest particles in the system. Loosely speaking, if the data behave like a power law of power p, the solution converges to the self-similar solution with the same power law. The notion "to behave like a power law" is made precise in [30], the technical term being that the data must be regularly varying with power p at the end of their support (for more details on regular varying functions see e.g. [1]).

To summarize, we have seen that the LSW theory has two short-comings. First, there is a significant discrepancy with experimental data and, second, there is only a weak selection of self-similar asymptotic states. Both of these shortcomings suggest that some important mechanisms have been neglected in the LSW model. Since it can be rigorously derived in the limit of vanishing volume fraction from the Mullins-Sekerka evolutions using averaging methods [28, 29, 31, 32], the model clearly represents the leading order theory. Thus, to

overcome the disadvantages of the LSW model, higher order effects have to be taken into account.

Higher order effects. There are at least two higher order effects of completely different nature, that have to be investigated. The first effect are screening induced fluctuations in the particles densities, the second encounters of particles.

By screening induced fluctuations we mean the following. Recall that in the derivation of the LSW mean-field model the crucial assumption was that one particle interacts with all the others only via a common mean-field. This neglects the screening effect, which is the same as in electrostatics, and implies that a particle is screened by the particles in its neighbourhood from the influence of the particles that are further away. As a consequence, larger particles that have been growing for a while, and more likely to be surrounded by smaller than average sized particles and can thus grow faster than predicted by the mean-field theory. Similarly, smaller particles shrink faster than predicted by the mean-field theory. In addition, the continuous vanishing of particles leads to the fact that the effective range of interactions of a particle gets larger in time, such that new particles enter the interaction radius. This type of noise could lead to an effective diffusion which again could act as a selection mechanism for self-similar solutions.

Encounters of particles happen on the other hand if two particles become close. Even though particles do essentially not move in space, the larger particles grow and finally come close to some other ones. These close particles eventually merge and form one larger particle. Since the largest particles dominate the long-time behaviour of the system, this effect has to examined closely.

In a first instance, one has to estimate the order of the size of these different correction terms. The screening effect can be most easily understood by referring to electrostatics. We briefly recall an argument that gives us the size of the correction coming from the screening effect. The parameters of a coarsening system are a number density n, the average radius $\langle R\rangle$ and the volume fraction of particles $\varepsilon \sim n\langle R\rangle^3 \ll 1$. Referring to electrostatics it can be seen that the screening length ξ, that is the effective range of particle interactions, scales as $\xi \sim \big(4\pi n\langle R\rangle\big)^{-1/2}$. Then the number of particles within the screening radius scales as $n\xi^3 \sim \varepsilon^{-1/2}$. As a consequence one expects that fluctuations are of a size of order $O(\varepsilon^{1/4})$ and the order of size of a correction term in the evolution equation for f should be $\varepsilon^{1/2}$. In comparison, the fraction of particles involved in collisions should scale as ε. Hence, one would expect that screening effects are more relevant, and perhaps for this reason this effect was investigated quite a lot in the applied literature (see e.g. [23, 24, 39–41]), whereas encounters are only studied rarely [21, 25].

A self-consistent derivation from the Mullins-Sekerka evolution that takes screening effects into account is given in [33]. The screening effects appear in an extension of the LSW model as a diffusion term in the size variable, that is the equation reads

$$\partial_t f + \partial_R\Big(\frac{1}{R^2}\Big(\frac{R}{\langle R\rangle} - 1\Big) f\Big) = \sqrt{\varepsilon}\partial_R\big(D(R)\partial_R f\big)\,, \tag{2.11}$$

where $\langle R\rangle$ denotes again the average radius and $D(R)$ is a complicated nonlocal term, which we do not give in detail here. A formal asymptotic analysis gives the existence of a unique self-similar solution to (2.11) with infinite support and exponential decay that furthermore predicts a correction to the average radius of the order $\varepsilon^{1/4}$.

In a next step one has to compare these predictions with the ones for a model that takes collisions of particles into account. Such an equation has already been suggested in [20]. It reads

$$\partial_t f + \partial_R\Big(\frac{1}{R^2}\Big(\frac{R}{\langle R\rangle} - 1)\Big) f\Big) = \varepsilon Q(f,f)(R)\,, \tag{2.12}$$

where Q is a quadratic coagulation term with a kernel that is additive in the rescaled volume variable. (We refer to the last section for a brief discussion of coagulation equations since Q looks awkward stated in the radius variable.)

The effect of the correction term on the right hand side of (2.12) has already been discussed on a formal level in the appendix of [20], but seems to have not been noticed much. At least it is not often discussed in papers that investigate the drawbacks and extension of the LSW model. The analysis suggests that there is a self-similar solution of (2.12) with infinite support and exponential decay that predicts a correction to the average growth rate of order $\frac{1}{|\ln\frac{1}{\varepsilon}|^2}$ which is indeed much larger than the correction given by (2.11). The analysis in [20] is based on a formal iteration scheme which was made fully rigorous in [15]. This result may on first glance seem very surprising given the discussion above. However, the explanation is quite intuitive. Equation (2.11) contains diffusive terms and changes the trajectory of each particle slightly. As we know from the discussion of the leading order LSW theory, the long-time behaviour is dominated by the largest particles in the system. There are only very few largest particles and so the effect of diffusion of these particles on the whole system is not very large. On the other hand, the coagulation term is a kinetic term and models that two particles can merge and form a large particle. In particular, two average sized particles can merge and form a large particle. However, there are many average sized particles in the system and thus this kinetic effect has a much larger influence on the long-time behaviour of the system than the diffusive type correction terms of (2.11).

3. Grain Growth

With grain growth one denotes coarsening of grains in a polycrystalline material. As described in the introduction, coarsening happens due to the desire of the system to reduce the surface energy of grains which is due to a mismatch of the orientation of the di fferent single crystal lattices.

We are interested in a kinetic description of a system of a large number of grains, which seems presently only possible in the significantly simpler case of two-dimensional grain growth. In the following we therefore restrict our considerations to two-dimensional networks of grain boundaries that meet in triple junctions. For simplicity we assume that these grains cover a finite rectangle and satisfy periodic boundary conditions on the boundary of this rectangle. As we have seen before, we need to identify the variables which characterize the grains. In the case of grains, we will need two quantities, the area a of a grain and its so-called topological class n, which is the number of neighbours, or equivalently the number of edges of the grain. In order to set up a kinetic mean-field model we need to derive how the area of a grain and the number of neighbours change in time. The difficulty lies in the fact that as compared to Ostwald Ripening there is no small parameter regime here that allows to derive such evolution laws in a certain limit.

In order to proceed one can however make the assumption that the surface energy density carried by the grain boundaries is constant, i.e. their energy is proportional to their length.

Furthermore one assumes that triple junctions have a large relative mobility such that one can assume that they adjust instantaneously to achieve local equilibrium of forces. As a consequence of these two assumptions the grain boundaries move according to the mean curvature flow while all angles at the triple junctions are $2\pi/3$. In this setting one can easily derive the celebrated von Neumann–Mullins law for the area $a(t)$ at time $t > 0$ of a single grain with n edges [27]:

$$\frac{d}{dt}a\,(t) = M\sigma\frac{\pi}{3}\,(n-6)\ . \tag{3.1}$$

Here M denotes the mobility of the grain boundaries and σ is the surface tension.

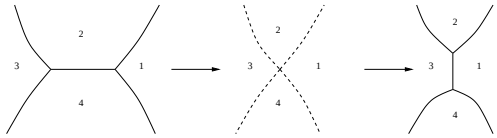


Figure 3.1. Neighbor switching

The mean curvature flow for such a network is well–defined [19, 22] until two vertices on a grain boundary collide, after which topological rearrangements may take place. If an edge vanishes, an unstable fourfold vertex is produced, which immediately splits up again such that two new vertices are connected by a new edge. As a consequence two neighbouring grains decrease their topological class (i.e., the number of edges), whereas the other two grains increase it (Fig. 3.1). Moreover, grains with topological class $2 \le n \le 5$ can vanish such that some vertices and edges disappear as illustrated in Figure 3.2. We now introduce

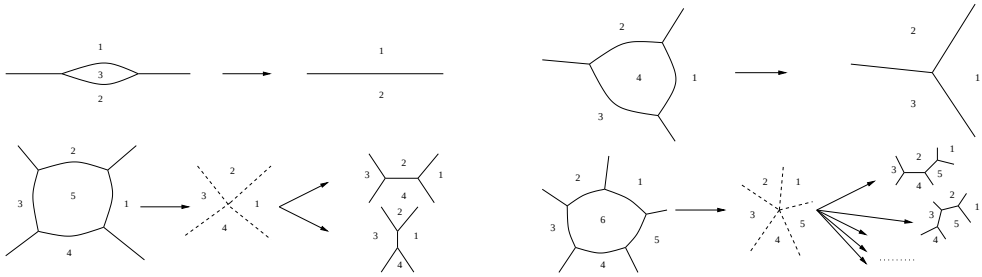


Figure 3.2. Grain vanishing

the number densities $f_n\,(a,t)$ of grains with topological class $n \ge 2$ and area $a \ge 0$ at time $t \ge 0$. As long as no topological rearrangements take place, the von Neumann–Mullins law (3.1) implies that f_n evolves according to

$$\partial_t f_n\,(a,t) + (n{-}6)\,\partial_a f_n\,(a,t) = 0\,.$$

Since the characteristics enter the domain $a > 0$ for $n > 6$, this equation must be supplemented with boundary conditions at $a = 0$ for $n > 6$, but one does not need any boundary condition for $n = 2, ..., 6$. It is reasonable to assume that no new grains are created during the coarsening process, which implies that

$$f_n\,(0,t) = 0 \qquad \text{for} \quad n \ge 7\,. \tag{3.2}$$

The key question now is how to incorporate topological changes in our model. Following Fradkov [11] one defines a 'collision' operator $\tilde{J}$ that couples the equations for different topological classes, that is one introduces topological fluxes η_n^+ and η_n^- that describe the flux from class n to $n+1$ and from n to $n-1$, respectively, and set

$$(\tilde{J}f)_n = \eta_{n-1}^+ + \eta_{n+1}^- - \eta_n^+ - \eta_n^-$$

with $\eta_1^+ = \eta_2^- = 0$ due to $n \geq 2$. In order to close the model one needs to express the fluxes η_n^+ and η_n^- in terms of f_n. This is only possible if one assumes that no correlations between the grains develop during the evolution. It is not at all clear, whether such an assumption is reasonable. However, it seems the only way to proceed to obtain a tractable kinetic model. Under this mean-field assumption, Fradkov [11] suggests that the fluxes are given by

$$\eta_n^+ = \Gamma\beta\, n f_n\,, \qquad \eta_n^- = \Gamma\,(\beta+1)\, n f_n, \tag{3.3}$$

where the coupling weight $\Gamma = \Gamma(f)$ describes the intensity of topological changes and depends on the complete state of the system in a self-consistent way, see (3.6) below. The free parameter β measures the ratio between switching and vanishing events and Fradkov et al [12] try to determine β by comparing the results of numerical simulations with experimental data. They suggest that β should roughly be 0.5.

Assumption (3.3) implies that the collision terms are given by $\tilde{J}f = \Gamma J f$ with

$$\begin{aligned} (Jf)_2 &= 3\,(\beta+1)\, f_3 - 2\beta f_2, \\ (Jf)_n &= (\beta+1)\,(n+1)\, f_{n+1} - (2\beta+1)\, n\, f_n + \beta\,(n-1)\, f_{n-1} \end{aligned} \tag{3.4}$$

for $2 < n < \infty$. Notice that this definition ensures the zero balance property

$$\sum_{n=2}^{\infty} (Jf)_n\,(a,\,t) = 0 \qquad \text{for all} \quad a,\, t > 0,$$

which reflects that the number of grains with given area does not change due to switching or vanishing events. To summarize, the complete kinetic model is given by

$$\partial_t f_n\,(a,t) + (n-6)\,\partial_a f_n\,(a,t) = \Gamma\big(f(t)\big)(Jf)_n\,(a,t)\,, \tag{3.5}$$

where $(a,t) \in (0,\infty)^2$ and $n \geq 2$ with boundary conditions (3.2) and $(Jf)_n$ given by (3.4). It remains to determine the coupling weightΓ in dependence of f. The key idea is to choose Γ such that the total area

$$A(t) = \sum_{n=2}^{\infty} Y_n(t) \qquad \text{with} \qquad Y_n(t) = \int_0^{\infty} a f_n(a,t)\, da$$

is conserved during the evolution. One easily checks that $dA/dt = P$, where P is the polyhedral defect defined by

$$P(t) = \sum_{n=2}^{\infty} (n-6) X_n(t) \qquad \text{with} \qquad X_n(t) = \int_0^{\infty} f_n(a,t)\, da.$$

The polyhedral formula $P = 0$ resembles Euler's formula for networks with triple junctions and states that the average number of neighbours per grain is 6. One now readily verifies that $dP/dt = 0$ holds if and only if

$$\Gamma\big(f(t)\big) = \frac{\sum_{n=2}^{5} (n-6)^2 \, f_n(0,t)}{\sum_{n=2}^{\infty} nX_n(t) - 2\,(\beta+1)\,X_2(t)}\,. \tag{3.6}$$

In particular, (3.6) guarantees the polyhedral formula as well as the conservation of area provided that the initial data satisfy $P = 0$. Well-posedness of the Fradkov model, both for $N < \infty$ and $N = \infty$, has been established in [14] for $\beta \in (0, 2)$. It is obvious from (3.6) that a main difficulty is to control the nonlocal quantity Γ. We also remark that a related simplified model has been considered in [6].

To study self-similar long-time behaviour of Fradkov's model by analytical tools seems to be very challenging. A first step that is already not easy is to prove the existence of a self-similar solution. This has been done in [16] and the main ideas of the proof are as follows.

Self-similar solutions to (3.5) take the form

$$f_n(a,t) = \frac{g_n(\xi)}{t^2}\,, \qquad \xi = \frac{a}{t} \geq 0\,,$$

where the sequence $g = (g_n)_{n\geq 2}$ of self-similar profiles satisfies

$$-2g_n - \big(\xi + 6 - n\big)g_n' = \Gamma\big(Jg\big)_n \tag{3.7}$$

for some positive constant Γ as well as the boundary conditions $g_n(0) = 0$ for $n > 6$.

The main mathematical difficulty in the existence proof for self-similar solutions is due to the fact that the ordinary differential equation (3.7) is singular at $\xi = n - 6$ and has different transport directions for $\xi < n - 6$ and $\xi > n - 6$. In [16] the existence of weak self-similar solutions is established, both for the system above and for a corresponding finite dimensional analogue, that is for the case that there is a maximal topological class $N < \infty$. Weak solution means that each function g_n satisfies

$$\int_0^\infty g_n\big((\xi + 6 - n)\phi' - \phi\big)\,d\xi + (6-n)_+\, g_n(0)\phi(0) = \Gamma(g)\int_0^\infty \big(Jg\big)_n \phi\,d\xi \tag{3.8}$$

for all smooth test functions ϕ with compact support in $[0,\,\infty)$.

The main result in [16] gives for $\beta \in (0, 2)$ the existence of a weak non-negative solution that decays fast in n and ξ in the sense that

$$\sum_{n=2}^{N} \Big(e^{\lambda n} X_n + \int_0^\infty e^{\lambda\xi} g_n(\xi)\,d\xi\Big) < \infty$$

for all $0 < \lambda < \ln\,(1 + 1/\beta)$, where $X_n = \int_0^\infty g_n(\xi)\,d\xi$.

The strategy for proving this result is inspired by the existence proof for self-similar solutions to coagulation equations in [10]. One first introduces a finite-dimensional dynamical model that can be regarded as a semi-discrete upwind scheme for (3.5) in self-similar variables, and which involves the discretization length $0 < \varepsilon \ll 1$. Standard results from the theory of

dynamical systems then imply the existence of nontrivial steady states for each sufficiently small ε. In a next step one shows that these steady states converge as $\varepsilon \to 0$ to a self-similar profile for the Fradkov model for $N < \infty$. To pass to the limit $N \to \infty$ one needs to establish exponential decay of X_n and uniform estimates for higher moments. The resulting tightness estimates allow then to obtain a solution for the infinite system.

The discrete scheme also gives naturally rise to a corresponding numerical algorithm that allows to study convergence to self-similar form by numerical simulations. These indicate that for any given set of parameters β, $A > 0$, and $6 < N < \infty$ there exists only one solution that is both self-similar and dynamically stable. However, it remains open whether there exist unstable self-similar solutions, and whether for $N = \infty$ there exist self-similar solutions that do not decay exponentially.

In conclusion, we have seen that the kinetic model for grain growth, despite the simplifications in the derivation, is still difficult to analyze. In addition to the crucial assumption that grains are uncorrelated, it relies heavily on the von-Neumann Mullins law, for which there is no three-dimensional analogue, and with this on the assumption that the surface energy of the grains are constant. Both assumptions are certainly very restrictive. Recent progress in numerical methods on the other hand, make numerical simulations of the full microscopic model competitive. For example, in [7, 8] develop numerical methods that can deal with a large number of grains (about 650000 in 2-d and 64000 in 3-d) and can also include surface energies that depend on the misorientation of the grains. The results of kinetic models should certainly be tested against the results from such simulations.

4. Connection with coagulation equations and conclusion

We briefly address here a related topic which is the analysis of self-similarity in Smoluchowski's mean-field model for coagulation. This equation has been derived by Smoluchowski in 1917 to qualitatively predict coagulation in a homogeneous colloidal gold solution. Since then this model has been used in a large variety of mass aggregation phenomena, for example in aerosol physics, polymerization, pattern formation in nanostructures, but also on very large scales in the clustering of stars.

In this model one considers a system of particles that are uniformly distributed in space and are characterized by their size $\xi \in (0, \infty)$, while $n(\xi, t)$ denotes the number density of particles of size ξ. The main assumptions in the model are that only binary coagulation is taken into account and that the rate at which two particles of size ξ and η coagulate is proportional to $n(\xi)n(\eta)$. The proportionality factor is given by a rate kernel $K(\xi, \eta)$ which is a symmetric, nonnegative function that represents all the microscopic details of the coagulation process. With these assumptions the rate equation for $n(\xi, t)$ becomes

$$\partial_t n(\xi,t) = \frac{1}{2}\int_0^\xi K(\xi-\eta,\eta)n(\xi-\eta,t)n(\eta,t)\,d\eta - n(\xi,t)\int_0^\infty K(\xi,\eta)n(\eta,t)\,d\eta\,. \quad (4.1)$$

Many different examples of kernels K can be found in the literature, but we mention here as a typical and important example only Smoluchowski's kernel

$$K(\xi,\eta) = \left(\xi^{1/3} + \eta^{1/3}\right)\left(\xi^{-1/3} + \eta^{-1/3}\right). \quad (4.2)$$

This kernel has been derived under the assumption that in $\mathbb{R}^3$ spherical clusters of diameter $\xi^{1/3}$ move independently by Brownian motion and coagulate quickly when they become

close. It is well known by now that this equation preserves the total mass $\int_0^\infty \xi n(\xi,t)\,d\xi$ if the kernel grows at most linearly, whereas gelation takes place, i.e. the loss of mass at finite time, if K grows faster than linearly. In both cases mass is shifted to larger and larger clusters as time proceeds and one expects that, for homogeneous kernels, this happens in a self-similar fashion. Thus, one is here also interested in finding self-similar solutions and characterize the large-time behaviour of solutions to (4.1) for given initial data. For the solvable kernels $K \equiv const.$, $K = \xi + \eta$ and $K = \xi\eta$, this issue is by now well-understood (see [26] and the references therein), but for all the other nonsolvable kernels, many questions are still unresolved. For kernels of homogeneity $\gamma < 1$ some progress has been made in recent years. The existence of self-similar solutions has been established [9, 10, 34], but uniqueness remains basically an open question even though recently a first such uniqueness result has been obtained for kernels that are close to constant [35]. Furthermore, also the question of dynamic stability of such solutions is still completely open. The difficulties in the analysis of this equation in self-similar variables are in principle the same as in the models that we discussed above. First, the model is nonlocal and no comparison principles are applicable. Furthermore, there is a competition between transport terms and a coagulation/collision term that has a diffusive character. The transport terms keep a memory of the initial data and the main question is whether the latter terms have enough mixing properties to drive the system nevertheless to a dynamic equilibrium.

Ackowledgments. I am very grateful to Bob Pego, Juan Velázquez, Felix Otto, Michael Herrmann, Philippe Laurençot and Joe Conlon for illuminating discussions and longstanding collaborations over the years on the subjects discussed here.

References

[1] N.H. Bingham, C.M. Goldie, and J.L. Teugels, *Regular variation. Paperback ed.*, Encyclopedia of Mathematics and its Applications, 27. Cambridge etc.: Cambridge University Press., 1989.

[2] L. C. Brown, *A new examination of classical Coarsening Theory*, Acta Metall. **37** (1989), 71–77.

[3] _____, *Reply to comments by Hillert, Hunderi, Ryum and Saetre on "A new examination of classical coarsening theory"*, Scripta Metall. **24** (1990), 963–966.

[4] _____, *Reply to comments by hoyt on "A new examination of classical coarsening theory"*, Scripta Metall. **24** (1990), 2231–2234.

[5] J. Carr and O. Penrose, *Asymptotic behaviour in a simplified Lifshitz–Slyozov equation*, Physica D **124** (1998), 166–176.

[6] A. Cohen, *A stochastic approach to coarsening of cellular networks*, Multiscale Model. Simul. **8** (2009/10), no. 2, 463–480.

[7] M. Elsey, S. Esedoglu, and P. Smereka, *Large scale simulation of normal grain growth via diffusion generated motion*, Proc. R. Soc. A **467:2126** (2011), 381–401.

[8] M. Elsey, S. Esedoglu, and P. Smereka, *Simulation of anisotropic grain growth: efficient algorithms and misorientation distributions*, Acta Materialia **61:6** (2013), 2033–2043.

[9] M. Escobedo, S. Mischler, and M. Rodriguez Ricard, *On self-similarity and stationary problem for fragmentation and coagulation models* , Ann. Inst. H. Poincaré Anal. Non Linéaire **22** (2005), no. 1, 99–125.

[10] N. Fournier and P. Laurençot, *Existence of self-similar solutions to Smoluchowski's coagulation equation*, Comm. Math. Phys. **256** (2005), no. 3, 589–609.

[11] V. E. Fradkov, *A theoretical investigation of two-dimensional grain growth in the 'gas' approximation*, Phil. Mag. Lett. **58** (1988), 271–275.

[12] V. E. Fradkov, D. G. Udler, and R. E. Kris, *Computer simulation of two–dimensional normal grain growth (the 'gas' approximation)*, Philos. Mag. Lett. **58** (1988), 277–283.

[13] B. Giron, B. Meerson, and P. V. Sasorov, *Weak selection and stability of localized distributions in Ostwald ripening*, Phys. Rev. E **58** (1998), 4213–6.

[14] R. Henseler, M. Herrmann, B. Niethammer, and Juan J.L. Velázquez, *A kinetic model for grain growth*, Kinetic and Related Models (KRM) **1** (2008), no. 4, 591 – 617.

[15] M. Herrmann, B. Niethammer, and J.J.L. Velázquez,*Self-similar solutions for the LSW model with encounters*, J. Differential Equations **247** (2009), 2282–2309.

[16] Michael Herrmann, Philippe Laurençot, and Barbara Niethammer, *Self-similar solutions to a kinetic model for grain growth*, J. Nonlinear Sci. **22** (2012), no. 3, 399–427.

[17] M. Hillert, O. Hunderi, and N. Ryum, *Instability of distribution functions in particle coarsening*, Scripta metall. **26** (1992), 1933–1938.

[18] M. Hillert, O. Hunderi, N. Ryum, and T. Saetre, *A comment on the Lifshitz-Slyozov-Wagner theory of particle coarsening*, Scripta metall. **23** (1989), 1979–1982.

[19] D. Kinderlehrer and C. Liu, *Evolution of grain boundaries*, Math. Models Methods Appl. Sci. **11** (2001), 713–729.

[20] I. M. Lifshitz and V. V. Slyozov, *The kinetics of precipitation from supersaturated solid solutions*, J. Phys. Chem. Solids **19** (1961), 35–50.

[21] A. Peleg P. V. Sasorov M. Conti, B. Meerson, *Phase ordering with a global conservation law: Ostwald ripening and coalescence*, Phys. Rev. E **65** (2002), 046117.

[22] C. Mantegazza, M. Novaga, and V. M. Tortorelli, *Motion by curvature of planar networks*, Ann. Sc. Norm. Super. Pisa Cl. Sci. **3** (2004), 235–324.

[23] M. Marder, *Correlations and Ostwald ripening*, Phys. Rev. A **36** (1987), 858–874.

[24] J. A. Marqusee and J. Ross, *Theory of Ostwald ripening: Competitive growth and its dependence on volume fraction*, J. Chem. Phys. **80** (1984), 536–543.

[25] L. Meli and P. F. Green, *Aggregation and coarsening of ligand-stabilized gold nanoparticles in poly(methyl methacrylate) thin films*, ACS NANO **2, 6** (2008), 1305–1312.

[26] G. Menon and R. L. Pego, *Approach to self-similarity in Smoluchowski's coagulation equations*, Comm. Pure Appl. Math. **57** (2004), no. 9, 1197–1232.

[27] W. W. Mullins, *Two–dimensional motion of idealized grain boundaries* , J. Appl. Phys. **27** (1956), 900–904.

[28] B. Niethammer, *Derivation of the LSW theory for Ostwald ripening by homogenization methods*, Arch. Rat. Mech. Anal. **147, 2** (1999), 119–178.

[29] B. Niethammer and F. Otto, *Ostwald Ripening: The screening length revisited*, Calc. Var. and PDE **13 1** (2001), 33–68.

[30] B. Niethammer and R. L. Pego, *Non–self–similar behavior in the LSW theory of Ostwald ripening*, J. Stat. Phys. **95, 5/6** (1999), 867–902.

[31] B. Niethammer and J. J. L. Velázquez, *Homogenization in coarsening systems I: deterministic case*, Math. Meth. Mod. Appl. Sc. **14,8** (2004), 1211–1233.

[32] ———, *Homogenization in coarsening systems II: stochastic case*, Math. Meth. Mod. Appl. Sc. **14,9** (2004), 1–24.

[33] ———, *On screening induced fluctuations in Ostwald ripening*, J. Stat. Phys. **130,3** (2008), 415–453.

[34] ———, *Self-similar solutions with fat tails for smoluchowski's coagulation equation with locally bounded kernels*, Comm. Math. Phys. **318** (2013), 505–532.

[35] ———, *Uniqueness of self-similar solutions to Smoluchowski's coagulation equations for kernels that are close to constant*, (2013), Preprint.

[36] P. W. Voorhees, *The theory of Ostwald ripening*, J. Stat. Phys. **38** (1985), 231–252.

[37] ———, *Ostwald ripening of two–phase mixtures*, Ann. Rev. Mater. Sc **22** (1992), 197–215.

[38] C. Wagner, *Theorie der Alterung von Niederschlägen durch Umlösen*, Z. Elektrochemie **65** (1961), 581–594.

[39] K. G. Wang and M. E. Glicksman, *Ostwald ripening in materials processing*, Materials Processing Handbook (J. R. Shackelford J. R. Groza, ed.), Taylor and Francis, 2007, pp. 5.1–20.

[40] K. G. Wang, M. E. Glicksman, and K. Rajan, *Modeling and simulation for phase coarsening: a comparison with experiment*, Phys. Rev. E (2004), 061507.

[41] J. H. Yao, K. R. Elder, H. Guo, and M. Grant, *Theory and simulation of Ostwald ripening*, Phys. Rev. B **47** (1993), 14110–14125.

Institut für Angewandte Mathematik, Universität Bonn, Endenicher Allee 60, 53115 Bonn, Germany
E-mail: niethammer@iam.uni-bonn.de

Computing global invariant manifolds: Techniques and applications

Hinke M. Osinga

Abstract. Global invariant manifolds play an important role in organising the behaviour of a dynamical system. Together with equilibria and periodic orbits, they form the so-called skeleton of the dynamics and offer geometric insight into how observed behaviour arises. In most cases, it is impossible to find invariant manifolds explicitly and numerical methods must be used to find accurate approximations. Developing such computational techniques is a challenge on its own and, to this date, the focus has primarily been on computing two-dimensional manifolds. Nevertheless, these computational efforts offer new insights that go far beyond a confirmation of the known theory. Furthermore, global invariant manifolds in dynamical systems theory not only explain asymptotic behaviour, but more recent developments show that they are equally useful for explaining short-term transient dynamics. This paper presents an overview of these more recent developments, in terms of novel computational methods, as well as applications that have stimulated recent advances in the field and highlighted the need for new mathematical theory.

Mathematics Subject Classification (2010). Primary 37C10; Secondary 37D10, 37C70, 65L10, 65P30.

Keywords. Dynamical systems, invariant manifold, boundary value problem, continuation techniques.

1. Introduction

Dynamical systems theory is very much characterised by its geometrical and topological aspects; classical textbooks, such as [6, 29, 33, 62, 63, 68], for example, rely on sketches to illustrate ideas. Therefore, it seems natural to have a computational toolbox that can produce numerical approximations to illustrate how this theory manifests itself in actual dynamical systems. The development of such a toolbox has proven to be a challenge in itself, which perhaps explains the apparent split of the field into those who use sketches and those who employ numerical computations; the two groups tend to interact too little. In fact, numerical computations are often used in realistic applications in collaboration with other scientists. There seems to exist a perception that this direction of research may lead to new numerical challenges, but does not contribute to the development of new theory, while theoreticians push the boundaries of dynamical systems and offer new insights via conjectures and then proofs. This paper aims to highlight how the development of dedicated computational methods arising from real applications can also lead to new dynamical systems theory. The focus

Proceedings of the International Congress of Mathematicians, Seoul, 2014

here will be on continuation methods for the computation of global invariant manifolds of vector fields.

Continuation methods for dynamical systems were designed for the bifurcation analysis of equilibria and periodic orbits. Pseudo-arclength continuation is used to track such invariant objects in a parameter [41]. Continuation of equilibria is relatively straightforward and involves finding an approximation to a uniquely defined parametrised solution family of an algebraic problem. The continuation of periodic orbits is already harder, because it requires solving a two-point boundary value problem (2PBVP) in conjunction with a suitable restriction to select a unique orbit from the infinite family of phase-shifted ones. The method of orthogonal collocation with piecewise polynomials [7, 10] is now widely adopted for this purpose, because it is very accurate and allows adaptive mesh selection; this particular solution method is implemented in the popular packages AUTO [16, 17], which is also part of the package XPPAUT [21], and MATCONT [15]. By extending the system to include suitable monitoring functions, the same approach can be used to continue codimension-one bifurcations in two parameters. In fact, the initiative behind the package MATCONT [15] aims to have implementations for the continuation of all codimension-one and -two bifurcations of equilibria and periodic orbits, both for continuous- and discrete-time deterministic systems [27, 46].

The continuation of periodic orbits is only one example of a 2PBVP set-up. Global invariant manifolds can also be formulated in terms of a 2PBVP. This idea has been applied to detect and continue homoclinic and heteroclinic bifurcations [36]. For example, the HOMCONT extension to AUTO can be used to compute such codimension-one bifurcations and determine the location of codimension-two points, such as homoclinic flip bifurcations [12]; these methods have also been developed for discrete-time systems [9], which is implemented for one-dimensional manifolds in the command-line version of MATCONT. Here, we apply the 2PBVP set-up in the context of computing two-dimensional global manifold of flows. We used AUTO to continue the 2PBVPs for the manifold computations in this paper. Four case studies illustrate the fruitful interplay between advancing the reach of the numerical methods and developing new dynamical systems theory.

This paper is organised as follows. In Section 2 we consider stable and unstable invariant manifolds, that is, manifolds that are globally invariant under the flow of the vector field and, either in forward or in backward time, converge to compact invariant objects, such as equilibria and periodic orbits. As specific examples, we consider the stable manifold of the origin in the Lorenz system in Section 2.1 and, in a more applied context in Section 2.2, the interpretation of a stable manifold as an isochrone for a particular phase point along a periodic orbit. In Section 3, we consider invariant manifolds as a tool to explain the effects of finite-time perturbations. In the example in Section 3.1, which is related to the notion of isochrones, we predict a delay or advance of the phase in response to a short-time perturbation. We then consider excitability in Section 3.2, and compute the excitability threshold in the context of a system for which no saddle equilibria or other saddle invariant manifolds are present. We conclude this review in Section 4 with a brief discussion that also mentions some directions of further research.

2. Stable and unstable manifolds

Stable and unstable manifolds of equilibria, periodic orbits, or other compact normally hyperbolic invariant manifolds of saddle type are an important part of the so-called *skeleton* of a dynamical system. While the attractors organise the eventual, asymptotic behaviour of the system, stable and unstable manifolds describe the global structure of the system, dictating which initial condition goes where, and in what manner.

To fix ideas and notation, let us restrict to vector fields from now on and consider global invariant manifolds of equilibria or periodic orbits. Recall that an equilibrium p is hyperbolic if all eigenvalues of the Jacobian matrix evaluated at p have non-zero real part; similarly, a periodic orbit Γ is hyperbolic if all Floquet multipliers of the linearisation have magnitudes different from 1, except for the Floquet multiplier associated with the direction tangent to Γ; we refer to [46] for details. The stable manifold of p or Γ, denoted $W^s(p)$ or $W^s(\Gamma)$, consists of all trajectories of the flow that converge to p or Γ in forward time; the unstable manifold of p or Γ, denoted $W^u(p)$ or $W^u(\Gamma)$, is its stable manifold when considering the time-reversed flow. The Stable Manifold Theorem [62] guarantees the existence of local (un)stable manifolds of hyperbolic equilibria and periodic orbits associated with their (un)stable eigenvalues or Floquet multipliers, and these manifolds can be extended globally by the flow in either forward or backward time. Furthermore, these manifolds are as smooth as the vector field itself, and they are tangent to the manifolds of the corresponding linearisation.

From these definitions, we deduce that a one-dimensional stable or unstable manifold of an equilibrium p of a vector field consists of two trajectories; each trajectory converges to p in forward or backward time, in a direction tangent to the eigenvector associated with the (strong) stable or unstable eigenvalue, such that the two trajectories together with p form a single smooth (immersed) manifold [62]. From a computational point of view, it is straightforward to compute such one-dimensional manifolds: by selecting an initial point along the appropriate eigenvector at a small distance from p, integration backward (for the stable manifold) or forward in time (for the unstable manifold) generates an orbit segment as an approximation of an arbitrarily long first piece of the manifold. Such an integration produces an ordered list of suitably distributed points on this first piece of the manifold, allowing for its straightforward visualisation as a smooth curve.

A two-dimensional (un)stable manifold, on the other hand, is a lot more difficult to compute and visualise. The challenge lies in the fact that the manifold is now a surface formed by a one-parameter family of trajectories. Hence, a computational method must include instructions how to generate a suitable mesh representation of this surface. Perhaps the simplest approach for designing an algorithm to compute two-dimensional (un)stable manifolds is to select (discretised) orbit segments from the one-parameter family that defines the manifold. Here, a first orbit segment is computed in the same way as for one-dimensional manifolds, by integration up to the time or length required. Continuation can then be used to follow this first orbit segment as its starting point is varied along a one-dimensional curve in the two-dimensional eigenspace; additional orbit segments are selected from the family as dictated by the spacing between them. This approach often requires a post-processing step of remeshing to visualise the surface. The complementary approach is to ignore the dynamics on the manifold and view it geometrically, for instance, as a family of geodesic level sets. In this case, the mesh is generated as a growing structure based on geometric features, and this aspect can be used for direct visualisation; the disadvantage is that the dynamics on the manifold may cause geometric obstructions, e.g., when there exists a connecting orbit

from one equilibrium or periodic orbit to another. We refer to the survey paper [45] for more details on these two (and other) approaches.

In the case studies presented here, we use both approaches, and each uses a formulation via two-point boundary value problems (2PBVP) that are solved by one-parameter continuation with the 2PBVP solver AUTO [16, 17]. We compute a finite set of (discretised) geodesic level sets with the algorithm from [42, 43] if we are interested in the two-dimensional manifold as a surface; this method generates a mesh with good geometric properties and allows for elaborate visualisation. We compute a one-parameter family of orbit segments [44, 45] if we are interested in how a manifold intersects another two- (or higher-)dimensional object, such as a plane or a sphere. Here, we compute the orbit segments up to this intersection and then consider and plot their end points; the orbit segments are selected based on a maximum distance between them, and so the end points give a good mesh representation of the intersection curves.

In the next sections we show how these computational methods can be employed to help understand the topological and geometric nature of the dynamics of a given system. In particular, they allow us to gain insights into different aspects of global dynamics, and we are even able to formulate precise conjectures based on our numerical findings.

2.1. The Lorenz manifold. As the leading example, we consider the stable manifold of the origin of the Lorenz equations. Recall that Lorenz introduced these equations as a much simplified model of convection in the atmosphere [48]. They take the form of three ordinary differential equations,

$$\left\{ \begin{array}{rcl} \dot{x} &=& \sigma\,(y-x), \\ \dot{y} &=& \varrho\,x - y - x\,z, \\ \dot{z} &=& x\,y - \beta\,z. \end{array} \right. \tag{2.1}$$

Lorenz used the classical values $\sigma = 10$, $\varrho = 28$ and $\beta = 8/3$ as representative parameters. The famous butterfly attractor is the associated globally attracting chaotic set. Note that the origin $\mathbf{0}$ is always an equilibrium of system (2.1), and it is of saddle type for the classical parameter values. There are two further, symmetrically-related equilibria, denoted $p^{\pm}$ that lie at the centres of the 'wings' of the butterfly attractor. The origin is hyperbolic with one unstable and two stable eigenvalues, which means that it has a one-dimensional unstable and a two-dimensional stable manifold. The equilibria $p^{\pm}$ each have a pair of complex conjugate unstable eigenvalues, with corresponding two-dimensional unstable manifolds, and one stable eigenvalue, with associated one-dimensional stable manifold. The two-dimensional stable manifold of the origin received its name Lorenz manifold in the survey paper [45] where all contributors used it as their test-case example. From a computational point of view, it is challenging to compute the Lorenz manifold, because there is an order of magnitude difference between the two stable eigenvalues. This means that, locally near the origin, a small disk will quickly transform into an elongated ellipse when carried by the flow backward in time. The nonlinear terms do not balance this effect, so that it is very hard to design algorithms that construct a high-quality mesh on the surface.

Figure 2.1 shows the Lorenz manifold $W^s(\mathbf{0})$ computed as a surface, that is, computed as a family of geodesic level sets [42, 43]. The outer boundary corresponds to the approximate geodesic level set at distance 162.5. The surface $W^s(\mathbf{0})$ is intersected with the plane $\Sigma_\varrho = \{z = \varrho - 1 = 27\}$, and the part of $W^s(\mathbf{0})$ that lies above Σ_ϱ, as well as Σ_ϱ itself, is rendered transparent. In this way, we can see the three equilibria $\mathbf{0}$ and $p^{\pm}$, with their one-dimensional manifolds: the unstable manifold $W^u(\mathbf{0})$ of $\mathbf{0}$ and the stable manifolds $W^s(p^{\pm})$

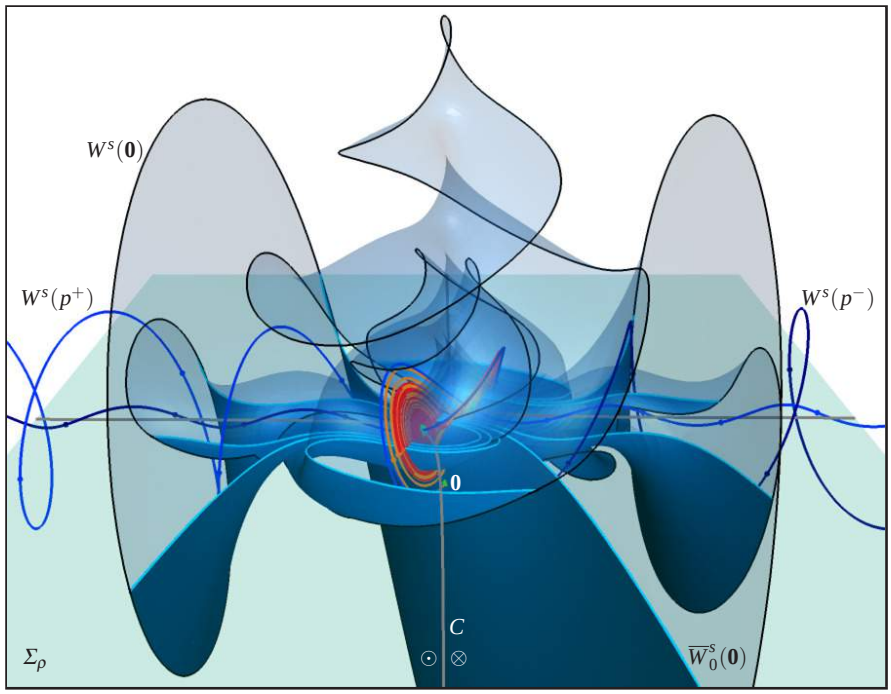


Figure 2.1. The Lorenz manifold $W^s(\mathbf{0})$, computed up to geodesic distance 162.5, and its intersection with the plane Σ_ϱ; the section Σ_ϱ and the part of $W^s(\mathbf{0})$ above it are rendered transparent. Also shown are the equilibria $\mathbf{0}$ and $p^\pm$, the one-dimensional manifolds $W^u(\mathbf{0})$ and $W^s(p^\pm)$, and the tangency locus C on Σ_ϱ. Reproduced from Osinga, Krauskopf, Hittmeyer, "Chaos and wild chaos in Lorenz-type systems," in Z. Al-Sharawi, J. Cushing and S. Elaydi (eds.) *19th Conference on Difference Equations and Applications* (in press), with permission from Springer-Verlag; see [59, Figure 4].

of $p^\pm$. The intersection curves and points of these manifolds with Σ_ϱ are also indicated. The plane Σ_ϱ is the Poincaré section that was used to analyse the nature of the dynamics on the attractor, which is believed to be the closure of $W^u(\mathbf{0})$. The return map is typically defined on the part in between $p^\pm$, where the flow points down from Σ_ϱ. The two hyperbolic curves denoted C separate this region from the regions where the flow points up; the flow is tangent to Σ_ϱ on C. The restriction of this return map to the Lorenz attractor can be approximated by a one-dimensional map, for which it is relatively straightforward to prove that it has chaotic dynamics [1, 30, 77]. The proof that the Lorenz attractor is indeed chaotic was completed only in 1999, and required computer assistance in the form of interval arithmetic [73, 76]. The reduction to a one-dimensional map requires the existence of a (one-dimensional) invariant foliation on Σ_ϱ that is transverse to the Lorenz attractor. We can see a few of the leaves in this foliation, namely, the intersection curves $\overline{W}^s(\mathbf{0}) := W^s(\mathbf{0}) \cap \Sigma_\varrho$; see [59] for more details.

The Lorenz manifold is a complicated surface. It cannot intersect (contain) the one-dimensional manifolds $W^s(p^\pm)$, and for the classical parameter values, it also does not intersect $W^u(\mathbf{0})$. In particular, due to the spiralling nature of $W^u(\mathbf{0})$ (and the attractor),

$W^s(\mathbf{0})$ winds in a helical manner around the z-axis, which is contained in $W^s(\mathbf{0})$, while additional helices are formed in symmetric pairs very close to but off the z-axis. At the same time, $W^s(\mathbf{0})$ spirals around $W^s(p^{\pm})$. Over the years, the challenge of computing the Lorenz manifold has shifted to the challenge of understanding its geometry. We view the Lorenz manifold as a key object for understanding how the chaotic dynamics manifests itself globally in the Lorenz system (2.1). Chaotic dynamics is characterised by the presence of sensitive dependence on initial conditions. Two nearby points on the Lorenz attractor quickly diverge under the flow; as a quantative measure, the signature or pattern of oscillations around p^+ and p^- will initially be identical, but after some time the two trajectories will move apart in such a way that the signature will be completely different. Switches between oscillations around p^+ and p^-, respectively, are organised by the close passage near $\mathbf{0}$. More precisely, $W^s(\mathbf{0})$ acts as a local separatrix between trajectories that continue oscillating around p^+, say, and those that switch to oscillating around p^-. Since the Lorenz attractor is a global attractor, any two points in phase space exhibit sensitive dependence on initial conditions, and this is organised globally by $W^s(\mathbf{0})$. This means that the global invariant manifold $W^s(\mathbf{0})$ separates any two points in $\mathbb{R}^3$ and is dense in $\mathbb{R}^3$.

It is mind-boggling to realise that such innocent-looking equations as the Lorenz system (2.1) give rise to a two-dimensional surface that lies dense in its three-dimensional phase space! This is an actual realised example of a space-filling surface. In order to visualise this topological property, and to study its characteristics further, we consider the intersection of $W^s(\mathbf{0})$ with a sphere S_R that is centred at the point $(0, 0, 27) \in \Sigma_\varrho$ on the z-axis (the mid-point on the line segment between $p^{\pm}$) and has large enough radius so that all bounded invariant objects are inside it; more precisely, we choose $R = 70.7099$, which is the distance from the centre of S_R to the second intersection point of the small-amplitude branch of $W^s(p^{\pm})$ with Σ_ϱ. Note that S_R is a compact surface so that any intersection curve with $W^s(\mathbf{0})$ must either be a closed curve or an arc with ends that accumulate on some sets, in this case the intersection points $W^s(p^{\pm}) \cap S_R$. Since $W^s(\mathbf{0})$ is dense in $\mathbb{R}^3$, the intersection curves in $\widehat{W}^s(\mathbf{0})$ must densely fill S_R.

Figure 2.2 shows $W^s(\mathbf{0})$ intersected with the sphere S_R. To highlight the situation on and inside S_R, only one half of $W^s(\mathbf{0})$ is shown, corresponding to the part that lies in the half space $\{y \geq 0\}$; the sphere S_R is rendered transparant. Many more curves in $\widehat{W}^s(\mathbf{0})$ are shown than those generated by the computed part of the surface $W^s(\mathbf{0})$. Indeed, the curves in $\widehat{W}^s(\mathbf{0})$ were computed directly, using the continuation of the family of trajectories that start on S_R and end on a small ellipse around $\mathbf{0}$ in the linear stable eigenspace of $\mathbf{0}$; the selected curves are associated with trajectories that satisfy these boundary conditions with a given maximal integration time [19]. The relatively large unfilled region on S_R shown in Figure 2.2 would be filled eventually, but only when an extremely large maximal integration time is used; two nearby points in these regions, while converging quickly to the Lorenz attractor, will take a comparatively large time to separate. Note the single curve that crosses through the middle of this region; it is the first intersection of $W^s(\mathbf{0})$ with S_R, that is, trajectories starting from points on this curve flow straight to $\mathbf{0}$ without excursions around p^+ or p^-. Hence, the unfilled region is directly related to the fact that trajectories on the Lorenz attractor visit a small neighbourhood of $\mathbf{0}$ far less frequently than similarly small neighbourhoods elsewhere on the Lorenz attractor [70, Appendix F]. Figure 2.2 also illustrates the structure of $\widehat{W}^s(\mathbf{0})$; the computed curves in $\widehat{W}^s(\mathbf{0})$ are the first of this set of curves that fills S_R densely, and they show that this process is taking place in a certain order associated with a Cantor set; see [19, 59] for more details.

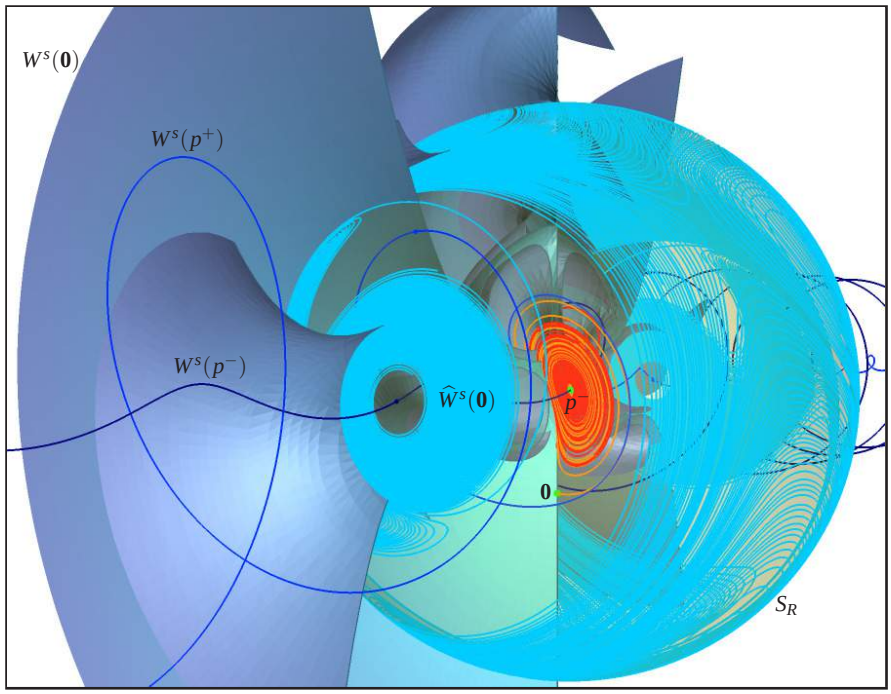


Figure 2.2. The Lorenz manifold $W^s(\mathbf{0})$ for $\varrho = 28$ intersecting the sphere S_R with $R = 70.7099$ in the set $\widehat{W}^s(\mathbf{0})$; also shown are the equilibria $\mathbf{0}$ and p^- and the one-dimensional manifolds $W^u(\mathbf{0})$ and $W^s(p^\pm)$. Reproduced from Osinga, Krauskopf, Hittmeyer, "Chaos and wild chaos in Lorenz-type systems," in Z. Al-Sharawi, J. Cushing and S. Elaydi (eds.) *19th Conference on Difference Equations and Applications* (in press), with permission from Springer-Verlag; see [59, Figure 2].

The study of the Lorenz manifold is ongoing, with a focus on the transitions that occur en route to chaos as a parameter is varied; often, ϱ is varied, which is proportional to the Rayleigh number of the convection [48]. For ϱ small enough, there is no chaotic dynamics. After a first homoclinic bifurcation, called the homoclinic explosion point, a so-called pre-turbulent regime is created, where a chaotic saddle is present; this first transition has been widely studied, for example, in [2, 18, 19, 38–40, 51–53, 64, 65, 70]. For details on the transition from pre-turbulent to turbulent dynamics, see also [18, 26, 80]; for more recent developments, see [13].

2.2. Isochrones. Isochrones were introduced in 1974 by Winfree [78] to characterise the behaviour of an oscillating system subjected to a brief external stimulus; the same external stimulus can have different effects depending on when it is applied. Such studies are useful, for example, to understand how signalling in neuronal networks is organised. Conceptually, the idea is very simple: the oscillations in the model are generated by an attracting periodic orbit Γ, which is typically assumed to be the only attractor in the system; any perturbation away from the periodic orbit, will result in a transient response that converges back to Γ, but perhaps with a different phase as before. The isochrones foliate the basin of attraction

of Γ in such a way that points on the same isochrone converge to Γ with the same phases. Guckenheimer [28] in a follow-up paper from 1975 explained that isochrones are nothing other than the pointwise stable manifolds of Γ. This means that each ischrone is invariant under the time-T map, where T is the period of Γ, and manifold theory can be used to show that isochrones must, therefore, be as smooth as the vector field itself and tangent to the linear stable eigenbundle of Γ [33].

From a geometric point of view, the isochrones form a nice manifold family that foliates the basin of attraction such that all isochrones accumulate on each other near the basin boundary. Winfree already realised this [25, 79], and studied the accumulation of one-dimensional isochrones in the two-dimensional FitzHugh–Nagumo system [24, 54] onto a repelling equilibrium enclosed by the attracting periodic orbit. Winfree expected to be able to compute the isochrones and visualise their geometry spiralling towards this repelling equilibrium, but to his surprise, he encountered serious numerical accuracy issues that could not be overcome at the time [79].

Isochrones have recently enjoyed a new surge of interest, fuelled in part by developments requiring controlled positioning onto specific isochrons. Numerous examples can be found in the context of biological applications, such as neuronal models, where the external stimulus represents a current injection coming from a large underlying neuronal network [23]. However, isochrones are also studied, for example, when regulating synchronisation of power networks that contain a large number of small energy generators, such as wind mills; see [47, 50, 60] for references. These important applications go hand in hand with a renewed interest in the development of appropriate numerical methods to compute isochrones [22, 31, 32, 37, 47, 49, 60, 69, 72]. In particular, we have overcome the accuracy issues reported by Winfree and are now able to compute the isochrones of the FitzHugh–Nagumo system reliably [47].

To illustrate some of these recent results, and discuss the difficulties encountered, we consider here a Hodgkin–Huxley model [35] that is reduced to the two-dimensional form studied in [60]. The model is described by the following system of two equations in terms of the membrane potential V and one of the gating variables n,

$$\left\{ \begin{array}{rcl} \dot{V} &=& -[I_{\rm Na} + I_{\rm K} + I_{\rm Leak}] + I_{\rm app}, \\ \dot{n} &=& \alpha_n(V)\,(1-n) - \beta_n(V)\,n. \end{array} \right. \tag{2.2}$$

Here, the different currents are given by

$$\begin{array}{rcl} I_{\rm Na} &=& g_{\rm Na}\,[m_\infty(V)]^3\,(0.8-n)\,(V-V_{\rm Na}), \\ I_{\rm K} &=& g_{\rm K}\,n^4\,(V-V_{\rm K}), \\ I_{\rm Leak} &=& g_{\rm L}(V-V_{\rm L}), \end{array}$$

and $I_{\rm app}$ is the applied current to stimulate the system so that an attracting periodic exists; we use $I_{\rm app} = 10$ throughout. The so-called quasi-steady-state function $m_\infty(V)$ is derived from the equilibrium assumption of a second gating variable m and is given by an equation of the same form as for n, that is,

$$m_\infty(V) = \frac{\alpha_m(V)}{\alpha_m(V) + \beta_m(V)}.$$

The functions $\alpha_j(V)$ and $\beta_j(V)$, with $j = n, m$ have the form

$$\alpha_j(V) = \frac{a_j(V+V_j)}{1-\exp[-(V+V_j)/k_j]} \quad \text{and} \quad \beta_j(V) = b_j \exp\left(\frac{-(V+E_j)}{\tau_j}\right).$$

$g_{\mathrm{Na}} = 120.0$	$g_{\mathrm{K}} = 36.0$	$g_{\mathrm{Leak}} = 0.3$
$V_{\mathrm{Na}} = 50.0$	$V_{\mathrm{K}} = -77.0$	$V_{\mathrm{L}} = -54.4$

$a_n = 0.01$,	$V_n = 55.0$,	$k_n = 10.0$	$b_n = 0.125$,	$E_n = 65.0$,	$\tau_n = 80.0$
$a_m = 0.1$,	$V_m = 55.0$,	$k_m = 10.0$	$b_m = 4.0$,	$E_m = 65.0$,	$\tau_m = 18.0$

Table 2.1. Parameters used in the two-dimensional reduced Hodgkin–Huxley model (2.2).

The particular constants used in this example are given in Table 2.1.

System (2.2) evolves on two different time scales; the membrane potential varies fast over a range of order $O(10^2)$, while n, which represents a fraction of open potassium channels, varies slowly over a unit range. While this time-scale separation is not made explicit in the model, one can see it in its spiking behaviour: the time series in V of the attracting periodic orbit Γ of this system has a long subthreshold plateau followed by a rapid large-amplitude spike. One main interest in such systems arises from the question whether it is possible to elicit a spike from the system via a small perturbation from an arbitrary point along the subthreshold plateau. It is generally believed that such a pertubation need only bring the system to a high enough level for V, the precise value of which is called the *spiking threshold*.

Figure 2.3 shows Γ together with 100 isochrones. The isochrones are distributed uniformly in time along Γ. This means that most isochrones are located on the subthreshold part, which is the lower, approximately horizontal segment of the closed (grey) curve in Figure 2.3(a). The isochrones are coloured according to a (cyan-to-magenta) colour gradient, starting from the maximal point on Γ (with respect to V), in the (clockwise) direction of the flow. Any perturbation away from Γ will land on a particular isochrone and relax back to Γ in phase with the point on Γ associated with this isochrone. The colour coding seems to reveal a clear spiking threshold, where all isochrones appear to align with each other. We focus on the situation near $n = 0.525$ and zoom into a neighbourhood of the perceived spiking threshold for this n-value, as shown in Figure 2.3(b). Here, we see that the isochrones do not merely align, but form a much more complicated structure, where each isochrone passes $n = 0.525$ several times while preserving its order in the foliation. This means that a perturbation close to the perceived spiking threshold could result in any arbitrary phase shift and the relationship between the size of the perturbation and the resulting phase shift, at least in this region of sensitivity, is highly nontrivial.

The characterisation of this stretched region of extreme phase sensitivity is related to the accumulation of isochrones near the basin boundary. Due to the two-dimensional nature of the flow, the periodic orbit Γ encloses an equilibrium at $(n, V) \approx (0.4026, -59.61)$, which is repelling. The enlargement in Figure 2.3(c) illustrates the intricate spiralling nature of the isochrones accumulating onto this equilibrium. The extreme phase sensitivity, not only near the equilibrium, is organised by the repelling slow manifold associated with the repelling branch of the cubic critical manifold; see [47, 60] for more details.

The computation of the isochrones uses a two-point boundary value set-up that is essentially the same as a stable-manifold calculation [47, 60]. We continue a one-parameter family of orbit segments with integration times equal to integer multiples of the period of Γ. By restricting one end point to a small interval along the linear stable eigendirection at a point $\gamma \in \Gamma$, the points at the other end of such a family of orbit segments forms the isochrone associated with γ. The resulting algorithm computes the isochrone as a curve parametrised by arclength and avoids the numerical accuracy issues reported by Winfree [79]. The con-

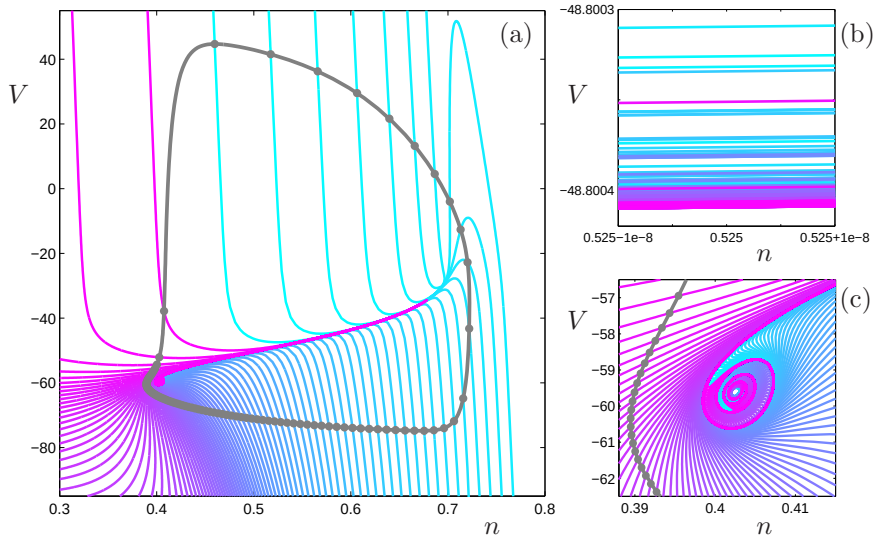


Figure 2.3. Extreme phase sensitivity near the excitability threshold in the reduced Hodgkin–Huxley model (2.2). Shown are the isochrones of 100 points along the periodic orbit (grey) that are distributed uniformly in time. Panel (b) shows the phase sensitivity in an enlargement near $n = 0.525$; and panel (c) illustrates how the isochrones organise the phase sensitivity near the equilibrium at $(n, V) \approx (0.4026, -59.61)$.

tinuation of the 2PBVP can trace the isochrone through regions of extreme phase sensitivity, because the entire orbit segments associated with ends points on different isochrones that are indistinguishable in this region, remain well separated.

3. Slow manifolds and transient effects

The example of the Hodgkin–Huxley model (2.2) in Section 2.2 illustrates that an excitability threshold can be much more complicated than generally assumed. Moreover, it highlights the need for a deeper mathematical understanding of bursting behaviour. The analysis of bursting goes back to the 1980s when Rinzel, at the 1986 ICM, proposed a simple approach to classifying bursting mechanisms in excitable systems [67]. Rinzel utilises the fact that excitable systems typically feature variables that evolve on rather different time scales. More precisely, the model can be written as

$$\begin{cases} \dot{x} = \quad f(x, y), \\ \dot{y} = \varepsilon\, g(x, y), \end{cases} \tag{3.1}$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, with $n, m \geq 1$. Here, $0 < \varepsilon \ll 1$ represents the single time-scale separation between y and x. If we take the singular limit $\varepsilon \to 0$ then y becomes a vector of parameters and the equation for x, called the fast subsystem, exhibits dynamics that depends on the choice for y. Rinzel discusses the case with $m = 1$ in detail. Bursting, or spiking, occurs when the one-parameter bifurcation diagram in y of the fast subsystem

exhibits hysteresis, and the y-nullcline is positioned such that the slow evolution of y causes an oscillation of y across this hysteresis regime. This idea of freezing the slow variable can even be used when ε is not explicitly present in the equations. For example, in the reduced Hodgkin–Huxley model (2.2), the variable V was found to be at least 100 times faster than n. Hence, one can view n as a parameter and analyse the one-dimensional fast subsystem given by the equation for V. Three equilibria co-exists for n approximately in the interval $[0.3085, 0.7072]$, both end points of which are fold points; the branches corresponding to the highest and lowest V-values are stable. Furthermore, n is decreasing on the lower branch and increasing on the upper branch in the hysteresis interval. One concludes that the full two-dimensional system exhibits a relaxation oscillation that traces the two branches of stable equilibria, interspersed by two (fast) jumps approximately at the fold points; the relaxation oscillation is the (gray) periodic orbit shown in Figure 2.3(a).

Different bursting patterns arise when there are additional bifurcations along the branches of equilibria. For example, multi-spike bursting oscillations arise when the upper branch includes a Hopf bifurcation, so that the fast subsystem exhibits periodic oscillations over a range of y-values; this case was already discussed in [67], but see also the example in the next section, where the fast subsystem undergoes a subcritical Hopf bifurcation, which gives rise to a family of unstable (saddle) periodic orbits, but nevertheless, generates a multi-spike burst. Bursting behaviour can also be organised by a slow-fast system with two or more slow variables; see [14] for a detailed discussion and literature overview.

The case studies presented in the following two sections are using the same ideas as introduced by Rinzel [66, 67], but utilise recent developments in manifold computations to enhance this approach and enlarge it applicability.

3.1. Predicting the phase response. In complete analogy to the two-dimensional reduced Hodgkin–Huxley model (2.2), we consider, here, the problem of phase resetting for a model of a pituitary cell. The model is four dimensional and uses the same Hodgkin–Huxley formalism as described in detail for system (2.2). One equation is for the membrane potential V, two are for channel gating variables n and m, and one is for calcium balance in the cell body:

$$\begin{cases} C_m \dot{V} &= -[I_{\mathrm{CaL}} + I_{\mathrm{CaT}} + I_{\mathrm{K}} + I_{\mathrm{KCa}} + I_{\mathrm{Leak}}] + I_{\mathrm{app}}, \\ \dot{n} &= \dfrac{n_\infty(V) - n}{\tau_n}, \\ \dot{m} &= \dfrac{m_\infty(V) - m}{\tau_m(V)}, \\ \dot{Ca} &= J_{\mathrm{exchange}} + f\,\beta\,(J_{\mathrm{influx}} - J_{\mathrm{efflux}}). \end{cases} \tag{3.2}$$

A full description of the model can be found in [71]; we only mention here that $I_{\mathrm{app}} = 0$ by default; it is only used for perturbing the spiking behaviour of the cell. Rather than eliciting a single spike, system (3.2) with $I_{\mathrm{app}} = 0$ exhibits a series of spikes during the active phase of the periodic orbit Γ. As for the reduced Hodgkin–Huxley model (2.2), most of the time is spent on a subthreshold plateau, and one is interested in understanding the response to perturbations away from this subthreshold segment of Γ. One particular difficulty with this model is to achieve an 'active' phase shift, in the sense that the perturbation brings the membrane potential up into the active phase and gives rise to a spike train before V drops back down to subthreshold levels.

System (3.2) has three different time scales: just as for the reduced Hodgkin–Huxley

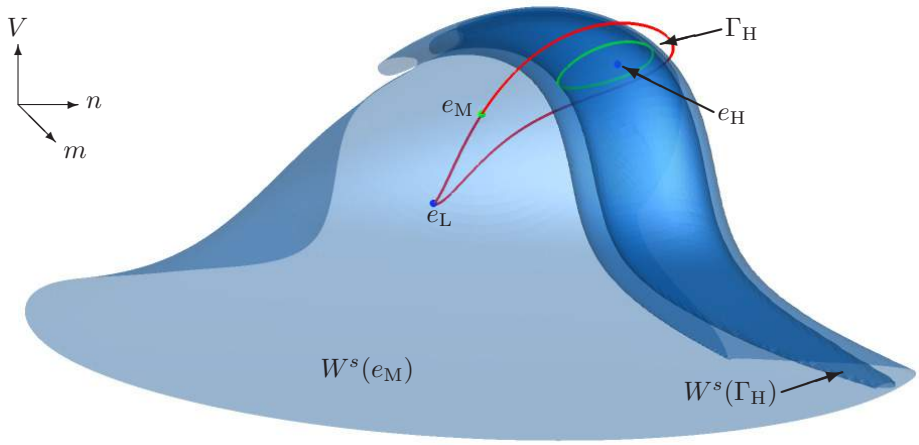


Figure 3.1. Stable manifolds of the equilibrium $e_{\rm M}$ and periodic orbit $\Gamma_{\rm H}$ of the fast subsystem of (3.2) with $Ca = 1$.

model (2.2), the membrane potential V varies on a much faster time scale than the two gating variables n and m. The calcium concentration varies even more slowly than the gating variables and it is this variable Ca that is singled out in the geometric singular perturbation theory, leaving a three-dimensional fast subsystem for analysis. The (V, n, m)-subsystem has two families of Ca-dependent stable equilibria, denoted $e_{\rm H}$ and $e_{\rm L}$ for the active and silent phases, respectively. The branch $e_{\rm L}$ exists only for large enough Ca, and coexists with a family $e_{\rm M}$ of saddle equilibria that meet at a fold. The branch $e_{\rm H}$ destabilises in a subcritical Hopf bifurcation for a Ca-value to the right of this fold point. Hence, there is a Ca-interval for which the two stable equilibria $e_{\rm H}$ and $e_{\rm L}$ coexist. The situation seems similar to the case discussed in Section 2.2, but the Hopf bifurcation gives rise to a family of saddle periodic orbits $\Gamma_{\rm H}$ that coexist with $e_{\rm H}$ and $e_{\rm L}$ for large enough Ca in the bistability interval.

We use the analysis of the fast subsystem to explain the difficulty in achieving an active phase shift. To this end, we focus on a single Ca-value, namely $Ca = 1$, for which all three equilibria as well as the saddle periodic orbit are present. A perturbation in the form of a current $I_{\rm app}$ is applied during the silent phase, such that $Ca = 1$, that is, (approximately) from the equilibrium $e_{\rm L}$. We assume that the transient effects caused by the perturbation are of such a short-time nature that Ca remains practically at 1. If this is indeed the case, then $I_{\rm app}$ must be such that $e_{\rm L}$, which for this new value of $I_{\rm app}$ is most certainly no longer an equilibrium, flows towards the basin of attraction of $e_{\rm H}$. Again, we assume that this transient motion is so fast that Ca hardly changes. As soon as the basin boundary is crossed, $I_{\rm app}$ can be switched off and we may assume that the dynamics will switch back to its unperturbed course with the required phase shift. Figure 3.1 shows the equilibria and periodic orbit of the fast subsystem for $Ca = 1$. Also shown are the two-dimensional stable manifolds $W^s(e_{\rm M})$ and $W^s(\Gamma_{\rm H})$ of $e_{\rm M}$ and $\Gamma_{\rm H}$, respectively. The manifolds $W^s(e_{\rm M})$ and $W^s(\Gamma_{\rm H})$ were computed with the same method described in Section 2.1. The basin boundary of $e_{\rm H}$ is the separatrix $W^s(\Gamma_{\rm H})$, but $W^s(e_{\rm M})$ also acts as a kind of separatrix, because a crossing of $W^s(e_{\rm M})$ leads to one or more spikes before relaxation back to $e_{\rm L}$.

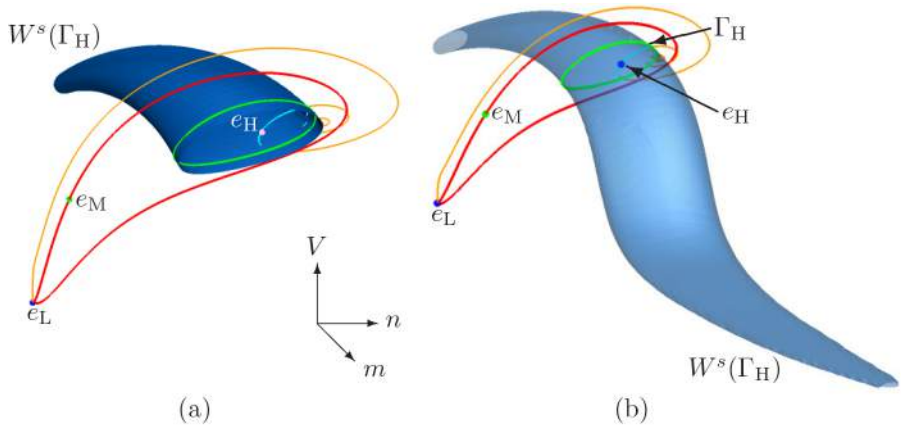


Figure 3.2. Starting from e_L, an applied current $I_{app} = 6.69$ results in two excursions inside the basin of attraction of e_H while spiralling towards an attractor outside this basin.

For $I_{app} > 0$ small enough, the fast subsystem has a similar set of three equilibria and one periodic orbit. Hence, for $I_{app} > 0$ small enough, the flow will simply push e_L to the corresponding (lower) stable equilibrium for the new value of I_{app}; this will not lead to an active phase shift. For $I_{app} > 0$ large enough, however, only one equilibrium exists, which can be associated with the active phase. For example, if $I_{app} = 6.69$, a unique attracting equilibrium exists near e_H. Unfortunately, this equilibrium lies outside the basin of attraction of e_H. This is the case for all values of I_{app} for which only one equilibrium exists. Figure 3.2 illustrates the possible transient behaviour while $I_{app} = 6.69$. The trajectory departs from e_L and spirals towards the attractor for this I_{app}-value. On its way, $W^s(\Gamma_H)$ is crossed four times, creating two short time windows in which the applied current could be reset to $I_{app} = 0$ and an active phase shift could possibly occur.

From this analysis we predict two successful perturbation protocols, both of which require holding I_{app} at a positive value for a certain (nontrivial) amount of time. Subsequent dynamic testing of these perturbation protocols for the full four-dimensional system indeed showed that an active phase shift can be achieved only for two particular segments in the silent phase. Perhaps more importantly, this research provided the precise ranges of values to use for I_{app} and the time duration befor reset to $I_{app} = 0$; until these results were known, researchers had been unable to find any kind of active phase reset for this type of pituitary cell model. We refer to [71] for more details.

It is interesting to note that the stable manifold of the coexisting saddle equilibrium e_M controls the number of spikes seen in a transient burst. The accumulation of $W^s(e_M)$ onto $W^s(\Gamma_H)$ occurs in the fast subsystem, but it is very similar to the isochrones accumulating onto a slow manifold, which occurs in the full system; for example, see the structure of the isochrones for the reduced Hodgkin–Huxley model in Section 2.2. As yet, there are no good methods available to compute higher-dimensional isochrones and the precise analogy remains a challenging area of research.

3.2. Excitability thresholds. The idea of using an applied current to elicit a spike or spike train from the model can be further refined to establish exactly how many spikes will be

generated after such a perturbation. In [56, 57] we considered a five-dimensional model that closely mimics the bursting behaviour of a pyramidal neurone in the so-called CA1 and CA3 regions of the hippocampus. Such CA1/CA3 cells are known to play an important role in the onset of Alzheimer's disease [5, 11, 55]. In experiments, these cells are subjected to a short current injection and the response of their membrane potential is recorded. A model for such a cell, constructed with the Hodgkin–Huxley formalism, offers insight into how the different currents bring about the various responses. Furthermore, the model can give a precise mathematical mechanism explaining how new spikes in the spike train are added when a parameter is varied.

The model combines equations for the membrane potential and four gating variables, corresponding to activation of slow inward and fast and slow outward currents, and inactivation of the slow inward current. Here, we consider only the model for a CA3 pyramidal neurone; the model for the CA1 neurone can be obtained by using a different set of parameters [55]. The parameters are such that the system is at its resting potential, which is an attracting equilibrium in the model; we refer to [56] for more details on the model equations. We study the transient response of this system when it is perturbed away from the stable equilibrium by an applied current of $20\,\mu\mathrm{A}/\mathrm{cm}^2$ for a duration of only $3\,\mathrm{ms}$. When the conductance parameter g_{SI} corresponding to the slow inward current is varied, this same short current-injection protocol leads to a variety of responses. More precisely, the strength and duration of the applied current is chosen such that, over a range of g_{SI}-values, the perturbation pushes the system past the top of a first spike; the difference between responses is characterised by what happens after the current injection, during the transient phase when the applied current is switched off and the system relaxes back to its stable equilibrium. Figure 3.3 shows three such responses, namely, for $g_{\mathrm{SI}} = 0.1$, for which the response immediately relaxes back to equilibrium, $g_{\mathrm{SI}} = 0.5$, for which the relaxation occurs via a non-monotonic route, and $g_{\mathrm{SI}} = 0.6\,\mathrm{mS}/\mathrm{cm}^2$, for which the response exhibits two further spikes before relaxation back to equilibrium.

The transformation from a single-spike to a three-spike response occurs via a spike-adding sequence, but the g_{SI}-interval of the two-spike response is very small and an example of such a response is not shown in Figure 3.3. In fact, experimental findings also report that it is difficult to obtain a two-spike response [11]. In order to investigate the mechanism underlying the spike-adding behaviour, at least from a mathematical point of view, we use geometric singular perturbation theory by utilising the different time scales in the model. Both the gating variables m_{SO} and h_{SI}, corresponding to activation of the slow outward current and inactivations of the slow inward current, respectively, are much slower than the other variables. Therefore, we consider the fast subsystem, represented by the membrane potential V, and the gating variables m_{SI} and m_{FO} corresponding to the slow inward and fast outward currents, respectively.

Since we now have two slow variables, the equilibria in this fast subsystem are organised in families that form surfaces in the five-dimensional phase space. In fact, they form a single folded sheet, if one allows h_{SI} to attain non-physical values. The lower segment (with respect to V) of this sheet consists of attracting equilibria, one of which corresponds to the stable equilibrium of the full five-dimensional system. The upper segment (with respect to V) is organised in much the same way as for the fast subsystem of (3.2) in Section 3.1: there exists a curve of subcritical Hopf bifurcations, which give rise to a two-parameter family of saddle periodic orbits. For the CA3 neurone model, this family of saddle periodic orbits undergoes a fold that stabilises the family before ending at a curve of homoclinic bifurcations. Figure 3.4

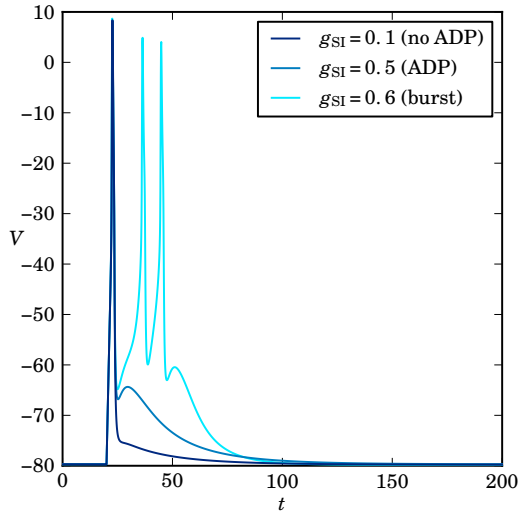


Figure 3.3. The same short current-injection protocol leads to different responses when a parameter is varied. Reproduced from Nowacki, Osinga, Tsaneva-Atanasova, "Dynamical systems analysis of spike-adding mechanisms in transient bursts," *Journal of Mathematical Neuroscience* **2** (2012): 7, with permission from Springer-Verlag; see [56, Figure 1].

shows these two-parameter families of equilibria and maxima and minima of the periodic orbits for $g_{\rm SI} = 0.5615$, which is a special value with respect to the behaviour of the full system, but representative for the geometric organisation of the equilibria and periodic orbits of the fast subsystem. The projection is onto $(h_{\rm SI}, m_{\rm SO}, V)$-space, showing V against the two slow variables $h_{\rm SI}$ and $m_{\rm SO}$. The surface of equilibria is labelled in segments according to the stability changes due to fold or Hopf bifurcations. The lower sheet is labelled S^a_1; past the first fold, — which occurs along a curve with $h_{\rm SI}$ outside its physical range and is not shown in Figure 3.4, — the equilibria are of saddle type and labelled S^r_1. There are two further folds that occur in quick succession, leading to an attracting segment S^a_2 and another saddle segment S^r_2. The upper fold (with respect to V) gives rise to a segment for which the equilibria have two unstable eigenvalues, and is labelled S^r_3; the upper attracting segment, on the other side of the Hopf curve, is labelled S^a_3. Similarly, the families of periodic orbits are denoted P^r and P^a.

Overlayed on the two-parameter families of equilibria are orbit segments of trajectories of the full five-dimensional system, starting from the point when the current injection has been switched off. From panels (a) to (f), the conductance $g_{\rm SI} \approx 0.5615$ is increasing, but only over an exponentially small interval; all $g_{\rm SI}$-values round to 0.5615. Figure 3.4 illustrates the significance of this value $g_{\rm SI} \approx 0.5615$, because in an exponentially small interval near this value, the orbit segment undergoes a dramatic transition that causes the creation of a new spike. While it is hard to see from such three-dimensional projections how this is organised in the five-dimensional phase space, Figure 3.4 gives a clear impression that the orbit segment tracks the unstable sheets S^r_1 and S^r_2 during the transition; we checked that this is indeed the case. A new spike is created when, at a special parameter value for $g_{\rm SI}$, the orbit

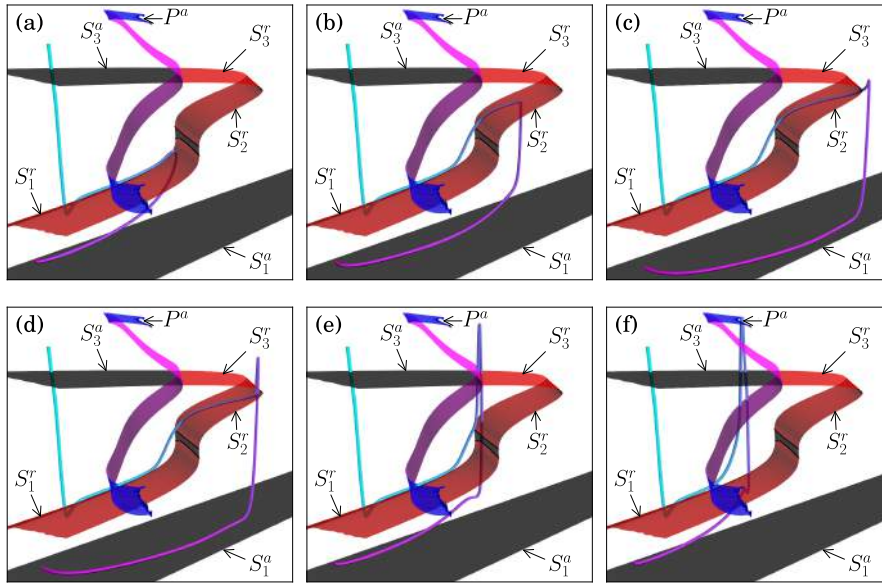


Figure 3.4. A spike-adding transition for the CA3 pyramidal neurone model with $g_{\rm SI} \approx 0.5615$ increasing over an exponentially small interval. Reproduced from Nowacki, Osinga, Tsaneva-Atanasova, "Dynamical systems analysis of spike-adding mechanisms in transient bursts," *Journal of Mathematical Neuroscience* **2** (2012): 7, with permission from Springer-Verlag; see [56, Figure 5].

segment does not immediately relax back to S_1^a, but is captured by the sheet S_1^r. At first, the orbit segment tracks S_1^r for only a short while before dropping down to S_1^a; see Figure 3.4(a). However, as $g_{\rm SI}$ increases, the orbit segment not only tracks S_1^r, but continues along S_2^r up to its fold with S_3^r before dropping back downd to S_1^a; see Figures 3.4(b) and (c). The transformation proceeds via the topological change that, after tracking S_1^r and S_2^r, the orbit segment jumps up before dropping down to S_1^a; see Figure 3.4(d). Subsequently, the tracking along S_1^r and S_2^r is gradually withdrawn, while the jump up develops into a real spike. We remark that the spike-adding transition for the CA3 neurone model is relatively complicated, involving two slow variables and a transition between two saddle-unstable sheets S_1^r and S_2^r. There features are important for the biology and help mimic precise details of the experimental results. However, the minimal ingredients for a spike-adding transition as illustrated in Figure 3.4 can be provided by a three-dimensional model with a single slow variable; see [61].

The spike-adding transition is initialised at the moment when the perturbation at the end of the short current injection is such that the orbit segment is captured by S_1^r. If we assume that the two slow variables $h_{\rm SI}$ and $m_{\rm SO}$ hardly change, we can illustrate this capture in $(m_{\rm SI}, m_{\rm FO}, V)$-space with respect to the fast subsystem. Figure 3.5 shows two views of the stable manifold of the saddle equilibrium $e_{\rm M}$ on S_1^r for the fast subsystem in $(m_{\rm SI}, m_{\rm FO}, V)$-space with $h_{\rm SI} = 0.6865$ and $m_{\rm SO} = 0.02534$; in both views, the vertical axis is V. The manifold $W^s(e_{\rm M})$ separates the basins of attraction of the two stable equilibria $e_{\rm L}$ on S_1^a and $e_{\rm H}$ on S_3^a; compare Figure 3.4. In the full five-dimensional phase space, $W^s(e_{\rm M})$ is

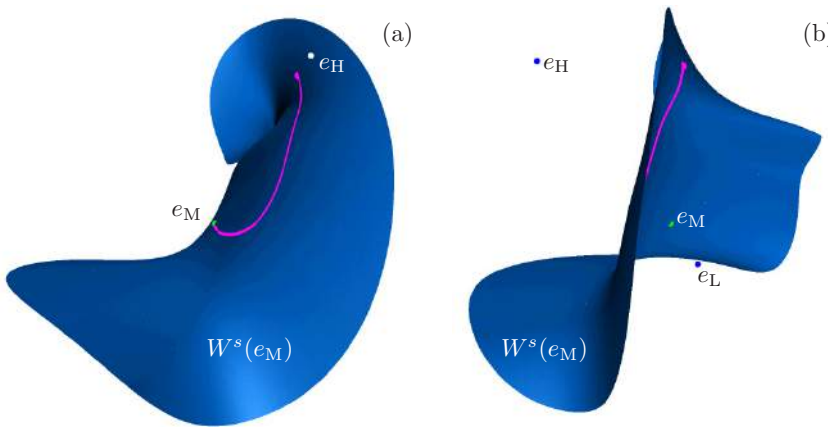


Figure 3.5. The stable manifold of the saddle equilibrium e_{M} on S_1^r with $h_{\mathrm{SI}} = 0.6865$ and $m_{\mathrm{SO}} = 0.02534$.

not a separatrix; it is not even an invariant manifold and e_{L}, e_{M} and e_{H} are not equilibria. We interpret Figure 3.5 in the following way. A spike-adding transition occurs when the parameter g_{SI} is such that the trajectory perturbed from the stable equilibrium of the full system lands exponentially close to $W^s(e_{\mathrm{M}})$ immediately after the $3\,\mathrm{ms}$ current injection. Here, $W^s(e_{\mathrm{M}})$ represents the stable manifold of the equilibrium e_{M} on S_1^r that corresponds to the (approximate) h_{SI}- and m_{SO}-values at the time immediately after the $3\,\mathrm{ms}$ current injection. As shown in Figure 3.5, the trajectory of the full five-dimensional system starts at a point near e_{H}, because the perturbation gave rise to a first spike. It lies (approximately) on $W^s(e_{\mathrm{M}})$ and, thus, converges to e_{M}. Since the fast directions dominate, h_{SI} and m_{SO} hardly change at first, and we can follow the convergence almost up to e_{M} in this 'frozen' image. Close to e_{M}, or more precisely, close to S_1^r, the slow dynamics dominates and the trajectory starts tracking S_1^r with h_{SI} and m_{SO} varying over a relatively large range; see Figure 3.4.

The excitability threshold in this system is not organised by the existence of a stable manifold in the full system, associated with a saddle equilibrium or other saddle-type invariant object. The role of the excitability threshold is taken over by unstable (saddle) slow manifolds that exist due to the presence of multiple time scales in the system. As argued here, the spike-adding dynamics is organised by the special events when a perturbation causes a shift exactly onto the stable manifold of a saddle slow manifold. One must be cautious here, because neither slow manifolds nor their stable manifolds are uniquely defined [14, 20]. In our case study, we consider the situation in the singular limit, for which the required stable manifold is uniquely defined, but for the full system, this means that the spike adding will be spread over an exponentially small parameter interval, during which the pertubation causes a shift onto stable manifolds of a family of saddle slow manifolds. The precise nature of such a transition, while observed numerically, has yet to be analysed in detail theoretically.

4. Conclusions

The case studies presented in this paper demonstrated that the continuation of two-point boundary value problems for the computation of global invariant manifolds is a powerful tool for the investigation of practical issues arising in applications, as well as questions in dynamical systems theory. In fact, these methods are so accurate that they allow for detailed quantitative predictions and the formulation of specific conjectures. Computations based on boundary-value-problem formulations can be used widely in dynamical systems; in particular, they are very well suited for the investigation of systems with multiple time scales. Moreover, they allow for a systematic investigation of transient phenomena.

We conclude this paper by mentioning a few directions of future research. In related and ongoing work, we consider the organisation of phase space near global bifurcations, including the Shilnikov bifurcation [4] and homoclinic flip bifurcations [3]. We also want to explore higher-dimensional systems, with a particular focus on hetero-dimensional cycles; an example with explicit equations of a system with hetero-dimensional cycles has only recently been found [81]. Such cycles are known to be related to the existence of wild chaos that can arise in vector fields of dimension at least four [8, 34, 74, 75]. We also continue our study of systems with multiple time scales and are particularly interested in interactions between slow manifolds and global invariant manifolds of such systems [14]. Furthermore, we would like to characterise the different mechanisms of spike adding in transient bursts [57, 61]. Finally, the computational approach to analyse transient bursts can also be employed in different applications. We are particularly interested in the stability analysis of a structure during an earthquake. The so-called failure boundary in this problem is similar to the excitability threshold studied in this paper. Initial computations that employ continuation of a two-point boundary value problem to find such failure boundaries directly, show that the boundary is formed in a complicated way, composed of piecewise-smooth segments from the solution family [58].

Acknowledgments. The research presented in this paper is the result of several collaborations and I would like to express my sincere gratitude to my coauthors. First and foremost, I thank Bernd Krauskopf for his continued enthusiasm and drive to work with me for almost two decades on many exciting areas of dynamical systems; I am particularly grateful for his detailed comments on a first draft of this paper. I also thank Eusebius Doedel, Arthur Sherman and Krasimira Tsaneva-Atanasova for our fruitful joint research activities that already span more than ten years as well. Bernd and I had the privilege to work with PhD students Pablo Aguirre, Jennifer Creaser, Mathieu Desroches, Peter Langfield, and Stefanie Hittmeyer on the global manifold computations discussed here, and Krasi and I enjoyed working with our PhD student Jakub Nowacki on transient dynamics.

References

[1] Afrajmovich, V. S., Bykov, V. V., and Sil′nikov, L. P., *The origin and structure of the Lorenz attractor*, Sov. Phys. Dokl. **22** (1977), 253–255; translation from Dokl. Akad. Nauk SSSR **234**(2) (1977), 336–339.

[2] Aguirre, P., Doedel, E. J., Krauskopf, B., and Osinga, H. M., *Investigating the conse-*

quences of global bifurcations for two-dimensional invariant manifolds of vector fields, Discr. Contin. Dynam. Syst. – Ser. A **29**(4) (2011), 1309–1344.

[3] Aguirre, P., Krauskopf, B., and Osinga, H. M., *Global invariant manifolds near homoclinic orbits to a real saddle: (non)orientability and flip bifurcation*, SIAM J. Appl. Dynam. Syst. **12**(4) (2013), 1803–1846.

[4] _____, *Global invariant manifolds near a Shilnikov homoclinic bifurcation*, J. Comput. Dynam. **1**(1) (2014), 1–38.

[5] Andersen, P., Morris, R., Amaral, D., Bliss, T., and O'Keefe, J., *The Hippocampus Book*. Oxford University Press, New York, 2007.

[6] Arnol′d, V.I., *Geometrical Methods in the Theory of Ordinary Differential Equations*. Springer-Verlag, Berlin, 2^{nd} edition, 1988.

[7] Ascher, U. M., Christiansen, J., and Russell, R. D., *Collocation software for boundary value ODEs*, ACM Trans. Math. Software **7**(2) (1981), 209–222.

[8] Bamón, R., Kiwi, J., and Rivera-Letelier, J., Wild Lorenz like attractors, arXiv:math/0508045 (2006); available at http://arxiv.org/abs/math/0508045/.

[9] Beyn, W.-J. and Kleinkauf, J.-M., *The numerical computation of homoclinic orbits for maps*, SIAM J. Numer. Anal. **34** (1997), 1207–1236.

[10] De Boor, C. and Swartz, B., *Collocation at Gaussian points*, SIAM J. Numer. Anal. **10**(4) (1973), 582–606.

[11] Brown, J. T. and Randall, A. D., *Activity-dependent depression of the spike after-depolarization generates long-lasting intrinsic plasticity in hippocampal CA3 pyramidal neurons*, J. Physiol. **587**(6) (2009), 1265–1281.

[12] Champneys, A. R., Kuznetsov, Yu. A., and Sandstede, B., *A numerical toolbox for homoclinic bifurcation analysis*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **6**(5) (1996), 867–887.

[13] Creaser, J. L., Krauskopf, B., and Osinga, H. M., α-flips in the Lorenz system, Preprint of The University of Auckland (2013).

[14] Desroches, M., Guckenheimer, J., Krauskopf, B., Kuehn, C., Osinga, H. M., and Wechselberger, M., *Mixed-mode oscillations with multiple time scales*, SIAM Review **54**(2) (2012), 211–288.

[15] Dhooge, A., Govaerts, W., Kuznetsov, and Yu. A., MATCONT*: A* MATLAB *package for numerical bifurcation analysis of ODEs*, ACM Trans. Math. Software **29**(2) (2003), 141–164.

[16] Doedel, E. J., AUTO*, a program for the automatic bifurcation analysis of autonomous systems*, Congr. Numer. **30** (1981), 265–384.

[17] _____, AUTO-07P*: Continuation and bifurcation software for ordinary differential equations*, with major contributions from Champneys, A. R., Fairgrieve, T. F., Kuznetsov, Yu. A., Oldeman, B. E., Paffenroth, R. C., Sandstede, B., Wang, X. J., Zhang, C. (2007); available at http://cmvl.cs.concordia.ca/auto/.

[18] Doedel, E. J., Krauskopf, B., and Osinga, H. M., *Global bifurcations of the Lorenz manifold*, Nonlinearity **19**(12) (2006), 2947–2972.

[19] ———, *Global invariant manifolds in the transition to preturbulence in the Lorenz system*, Indag. Math. (N.S.) **22**(3–4) (2011), 222–240.

[20] Dumortier, F. and Roussarie, R., *Canard cycles and center manifolds*, Mem. Amer. Math. Soc. **121**, Providence, RI, 1996; with an appendix by Cheng Zhi Li.

[21] Ermentrout, G. B., *Simulating, Analyzing, and Animating Dynamical Systems* (A Guide to XPPAUT for Researchers and Students), SIAM, Philadelphia, 2002.

[22] Ermentrout, G. B., Glass, L., and Oldeman, B. E., *The shape of phase-resetting curves in oscillators with a saddle node on an invariant circle bifurcation*, Neural Computation **24**(12) (2012), 3111–3125.

[23] Ermentrout, G. B. and Terman, D. H., *Mathematical Foundations of Neuroscience*, Springer-Verlag, New York, 2010.

[24] FitzHugh, R., *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J. **1**(6) (1961), 445–466.

[25] Glass, L. and Winfree, A. T., *Discontinuities in phase-resetting experiments*, Amer. J. Physiol.-Regul., Integr. Compar. Physiol. **246**(2) (1984), R251–R258.

[26] Glendinning, P. and Sparrow, C., *Local and global behavior near homoclinic orbits*, J. Statist. Phys. **35** (1984), 645–696.

[27] Govaerts, W., Kuznetsov, and Yu. A., *Interactive continuation tools*, in Krauskopf, B., Osinga, H. M., Galán-Vioque, J. (eds.) Numerical Continuation Methods for Dynamical Systems, pp. 51–75. Springer-Verlag, New York, 2007.

[28] Guckenheimer, J., *Isochrons and phaseless sets*, J. Math. Biol. **1**(3) (1975), 259–273.

[29] Guckenheimer, J. and Holmes, P., *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, 2^{nd} edition, 1986.

[30] Guckenheimer, J. and Williams, R. F., *Structural stability of Lorenz attractors*, Publ. Math. IHES **50** (1979), 59–72.

[31] Guillamon, A. and Huguet, G., *A computational and geometric approach to phase resetting curves and surfaces*, SIAM J. Appl. Dynam. Syst. **8**(3) (2009), 1005–1042.

[32] Gutkin, B. S., Ermentrout, G. B., and Reyes, A. D., *Phase-response curves give the responses of neurons to transient inputs*, J. Neurophysiology **94**(2) (2005), 1623–1635.

[33] Hirsch, M. W., Pugh, C. C., and Shub, M., *Invariant Manifolds*. Lecture Notes in Math. **583**, Springer-Verlag, New York, 1977.

[34] Hittmeyer, S., Krauskopf, B., and Osinga, H. M., *Interacting global invariant sets in a planar map model of wild chaos*, SIAM J. Appl. Dynam. Syst. **12**(3) (2013), 1280–1329.

[35] Hodgkin, A. L. and Huxley, A. F., *A quantitive description of membrane current and its application to conduction and excitation in nerve*, J. Physiol. **117**(4) (1952), 500–544.

[36] Homburg, A. J. and Sandstede, B., *Homoclinic and heteroclinic bifurcations in vector fields*, in Broer, H. W., Takens, F., Hasselblatt, B. (eds.), Handbook of Dynamical Systems Vol. **3**, pp. 379–524. North-Holland, Amsterdam, 2010.

[37] Huguet, G. and de la Llave, R., *Computation of limit cycles and their isochrons: Fast algorithms and their convergence*, SIAM J. Appl. Dynam. Syst. **12**(4) (2013), 1763–1802.

[38] Jackson, E. A., *The Lorenz system: I. The global structure of its stable manifolds*, Physica Scripta **32**(5) (1985), 469–475.

[39] ———, *The Lorenz system: II. The homoclinic convolution of the stable manifolds*, Physica Scripta **32**(5) (1985), 476–481.

[40] Kaplan, J. L. and Yorke, J. A., *Preturbulence: A regime observed in a fluid flow model of Lorenz*, Commun. Math. Phys. **67** (1979), 93–108.

[41] Keller, H. B., *Numerical solutions of bifurcation and nonlinear eigenvalue problems*, in Rabinowitz, P. H. (ed.) Applications of Bifurcation Theory, pp 359–384. Academic Press, New York, 1977.

[42] Krauskopf, B. and Osinga, H.M., *Two-dimensional global manifolds of vector fields*, CHAOS **9**(3) (1999), 768–774.

[43] ———, *Computing geodesic level sets on global (un)stable manifolds of vector fields*, SIAM J. Appl. Dynam. Sys. **2**(4) (2003), 546–569.

[44] ———, *Computing invariant manifolds via the continuation of orbit segments*, in Krauskopf, B., Osinga, H. M., Galán-Vioque, J. (eds.) Numerical Continuation Methods for Dynamical Systems, pp 117–154. Springer-Verlag, New York, 2007.

[45] Krauskopf, B., Osinga, H.M., Doedel, E.J., Henderson, M. E., Guckenheimer, J., Vladimirsky, A., Dellnitz, M., and Junge, O., *A survey of methods for computing (un)stable manifolds of vector fields*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **15**(3) (2005), 763–791.

[46] Kuznetsov, Yu. A., *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 2nd edition, 1998.

[47] Langfield, P., Krauskopf, B., and Osinga, H. M., *Solving Winfree's puzzle: the isochrons in the FitzHugh-Nagumo model*, Chaos **24**(1) (2014), 013131.

[48] Lorenz, E. N., *Deterministic nonperiodic flows*, J. Atmosph. Sci. **20** (1963), 130–141.

[49] Mauroy, A., and Mezić, I., *On the use of Fourier averages to compute the global isochrons of (quasi) periodic dynamics*, Chaos **22**(3) (2012), 033112.

[50] Mauroy, A., Mezić, I., and Moehlis, J., *Isostables, isochrons, and Koopman spectrum for the action-angle representation of stable fixed point dynamics*, Physica D **261** (2013), 19–30.

[51] Mischaikow, K. and Mrozek, M., *Chaos in the Lorenz equations: A computer assisted proof*, Bull. Amer. Math. Soc. **32**(1) (1995), 66–72.

[52] _____, *Chaos in the Lorenz equations: A computer assisted proof part II: Details*, Math. Comput. **67**(223) (1998), 1023–1046.

[53] Mischaikow, K., Mrozek, M., and Szymczak, A., *Chaos in the Lorenz equations: A computer assisted proof part III: Classical parameter values*, J. Diff. Equations **169** (2001), 17–56.

[54] Nagumo, J. S., Arimoto, S., and Yoshizawa, S., *An active pulse transmission line simulating nerve axon*, Proc. Inst. Radio Engineers **50** (1962), 2061–2070.

[55] Nowacki, J., Osinga, H. M., Brown, J. T., Randall, A. D., and Tsaneva-Atanasova, K. T., *A unified model of CA1/3 pyramidal cells: An investigation into excitability*, Progr. Biophys. Molec. Biol. **105**(1-2) (2011), 34–48.

[56] Nowacki, J., Osinga, H. M., and Tsaneva-Atanasova, K. T., *Dynamical systems analysis of spike-adding mechanisms in transient bursts*, J. Math. Neurosci. **2** (2012), 7.

[57] _____, *Continuation-based numerical detection of after-depolarisation and spike-adding threshold*, Neural Computation **25**(4) (2013), 877–900.

[58] Osinga, H. M., *Computing failure boundaries by continuation of a two-point boundary value problem*, in Proc. Ninth International Conference on Structural Dynamics, Porto, Portugal (in press).

[59] Osinga, H.M., Krauskopf, B., and Hittmeyer, S., *Chaos and wild chaos in Lorenz-type systems*, in Al-Sharawi, Z., Cushing, J., Elaydi, S. (eds.) 19th International Conference on Difference Equations and Applications, Springer-Verlag, New York (in press).

[60] Osinga, H. M. and Moehlis, J., *Continuation-based computation of global isochrons*, SIAM J. Appl. Dynam. Syst. **9**(4) (2010), 1201–1228.

[61] Osinga, H. M. and Tsaneva-Atanasova, K. T., *Geometric analysis of transient bursts*, Chaos **23**(4) (2013), 046107.

[62] Palis, J. and de Melo, W., *Geometric Theory of Dynamical Systems*, Springer-Verlag, New York, 1982.

[63] Palis, J. and Takens, F., *Hyperbolicity & Sensitive Chaotic Dynamics at Homoclinic Bifurcations*, Cambridge University Press, Cambridge, 1993.

[64] Perelló, C., *Intertwining invariant manifolds and Lorenz attractor*, in Global Theory of Dynamical Systems, pp. 375–378. Proc. Internat. Conf., Northwestern Univ., Evanston, IL, Lecture Notes in Math. **819**, Springer-Verlag, Berlin, 1979.

[65] Rand, D., *The topological classification of Lorenz attractors*, Math. Proc. Cambridge Philos. Soc. **83** (1978), 451–460.

[66] Rinzel, J., *Bursting oscillations in an excitable membrane model*, in Sleeman, B. D., Jarvis, R. J. (eds.) Ordinary and Partial Differential Equations (Dundee, 1984), pp. 304–316. Lecture Notes in Math. **1151**, Springer-Verlag, New York, 1985.

[67] Rinzel, J. *A formal classification of bursting mechanisms in excitable systems*, in Gleason, A. M. (ed.) Proc. Int. Congress Math., Berkeley 1986, Vol. 1, 2, pp. 1578–1593. Amer. Math. Soc., Providence, RI, 1987; also (with slight differences) in Teramoto, E., Yamaguti, M. (eds.) *Mathematical Topics in Population Biology, Morphogenesis and Neuroscience*, pp. 267–281. Lecture Notes in Biomath. **71**, Springer-Verlag, Berlin, 1987.

[68] Robinson, C., *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press LLC, Boca Raton, FL, 2^{nd} edition, 1999.

[69] Sherwood, W. E. and Guckenheimer, J., *Dissecting the phase response of a model bursting neuron*, SIAM J. Appl. Dynam. Syst. **9**(3) (2010), 659–703.

[70] Sparrow, C., *The Lorenz Equations: Bifurcations, Chaos and Strange Attractors*, Appl. Math. Sci. No. **41**, Springer-Verlag, New York, 1982.

[71] Stern, J. V., Osinga, H. M., LeBeau, A., and Sherman, A., *Resetting behavior in a model of bursting in secretory pituitary cells: Distinguishing plateaus from pseudo-plateaus*, Bull. Math. Biol. **70**(1) (2008), 68–88.

[72] Takeshita, D. and Feres, R., *Higher order approximation of isochrons*, Nonlinearity **23**(6) (2010), 1303–1323.

[73] Tucker, W., *The Lorenz attractor exists*, C. R. Acad. Sci. Paris Sér. I Math. **328**(12) (1999), 1197–1202.

[74] Turaev, D. V. and Shilnikov, L. P., *An example of a wild strange attractor*, Mat. Sb. **189** (1998), 291–314.

[75] ———, *Pseudo-hyperbolicity and the problem on periodic perturbations of Lorenz-like attractors*, Russian Dokl. Math. **77** (2008), 17–21.

[76] Viana, M., *What's new on Lorenz strange attractors?*, Math. Intell. **22**(3) (2000), 6–19.

[77] Williams, R. F., *The structure of Lorenz attractors*, Publ. Math. IHES **50** (1979), 101–152.

[78] Winfree, A. T., *Patterns of phase compromise in biological cycles*, J. Math. Biol. **1**(1) (1974), 73–93.

[79] ———, *The Geometry of Biological Time*, Interdisc. Appl. Math. **12**, 2^{nd} edition, Springer-Verlag, New York, 2001.

[80] Yorke, J. A. and Yorke, E. D., *Metastable chaos: The transition to sustained chaotic behavior in the Lorenz model*, J. Stat. Phys. **21** (1979), 263–277.

[81] Zhang, W., Krauskopf, B., and Kirk, V., *How to find a codimension-one heteroclinic cycle between two periodic orbits*, Discr. Contin. Dynam. Syst. – Ser. A **32**(8) (2012), 2825–2851.

Department of Mathematics, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

E-mail: h.m.osinga@auckland.ac.nz

Numerical approximation of variational inequalities arising in elastoplasticity

B. Daya Reddy

Abstract. Mathematical models of many classes of nonsmoooth problems in mechanics take the form of variational inequalities. Elastoplasticity, which is a theory of solids that exhibit path-dependent and irreversible behaviour, yields a variational inequality that is not of standard elliptic or parabolic type. Properties of the corresponding abstract problem are reviewed, as are the conditions under which fully discrete approximations converge. A solution algorithm, motivated by the predictor-corrector algorithms that are common in elastoplastic problems, is constructed for the abstract problem and shown to converge.

Mathematics Subject Classification (2010). Primary 65M60, 65M15, 74C05; Secondary 65N30.

Keywords. Elastoplasticity, variational inequalities, finite elements, algorithms, convergence, predictor-corrector schemes.

1. Introduction

Mathematical models of a large class of problems in solid and fluid mechanics take the form of systems of partial differential equations, in space and time. For example, the equation

$$\rho \frac{\partial^2 \boldsymbol{u}}{\partial t^2} - \bar{\Delta} \boldsymbol{u} = \boldsymbol{f} \tag{1.1}$$

describes the motion of an isotropic linear elastic solid. For a problem posed on a domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ and on a time interval $[0, T]$, $\boldsymbol{u} : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ is the displacement vector, $\boldsymbol{f}$ is a prescribed body force, ρ is the mass density of the body, and

$$\bar{\Delta} \boldsymbol{u} := (\lambda + \mu) \nabla \operatorname{div} \boldsymbol{u} + \mu \operatorname{div} \nabla \boldsymbol{u} \tag{1.2}$$

is the Lamé operator, with λ and μ being strictly positive material scalars. A complete description of the problem requires in addition a set of boundary and initial conditions. In the event that the data does not depend on time, (1.1) becomes the equilibrium equation

$$-\bar{\Delta} \boldsymbol{u} = \boldsymbol{f} \,. \tag{1.3}$$

Initial-boundary value or boundary value problems of this kind may be formulated alternatively in weak or variational form. In addition to providing a useful setting for establishing

Proceedings of the International Congress of Mathematicians, Seoul, 2014

well-posedness, the weak formulation also serves as the starting point for obtaining approximate solutions using the Galerkin finite element method. Consider for example the boundary value problem (1.3) with a homogeneous boundary condition: that is, $\boldsymbol{u} = \boldsymbol{0}$ on the boundary $\partial\Omega$. Setting $V = [H_0^1(\Omega)]^d$ in which $H_0^1(\Omega) = \{v \in L^2(\Omega), \partial v/\partial x_i \in L^2(\Omega), v = 0 \text{ on } \partial\Omega\}$ is the Sobolev space of functions with zero trace on the boundary, the weak form of the boundary value problem corresponding to (1.3) is that of finding $\boldsymbol{u} \in V$ that satisfies

$$a(\boldsymbol{u}, \boldsymbol{v}) = \langle \ell, \boldsymbol{v} \rangle \qquad \forall \boldsymbol{v} \in V\,. \tag{1.4}$$

Here $a(\cdot\cdot) : V \times V \to \mathbb{R}$ is a bilinear form and $\langle \cdot, \cdot \rangle : V' \times V \to \mathbb{R}$ denotes duality pairing between members of the topological dual V' and V. The bilinear form a and linear functional ℓ are defined by

$$a(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \Big[\lambda(\operatorname{div} \boldsymbol{u})(\operatorname{div} \boldsymbol{v}) + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\Big]\, dx\,, \tag{1.5a}$$

$$\ell(\boldsymbol{v}) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx\,, \tag{1.5b}$$

with

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) := \tfrac{1}{2}(\nabla \boldsymbol{u} + [\nabla \boldsymbol{u}]^T) \tag{1.6}$$

being the symmetric gradient of $\boldsymbol{u}$, or strain tensor: in component form $\epsilon_{ij}(\boldsymbol{u}) = \frac{1}{2}(\partial u_i/\partial x_j + \partial u_j/\partial x_i)$. Problem (1.4) has a unique solution given that there are positive constants C and α such that a is continuous: $|a(\boldsymbol{u}, \boldsymbol{v})| \leq C\|\boldsymbol{u}\|_V \|\boldsymbol{v}\|_V$, and V-ellliptic: $(\boldsymbol{v}, \boldsymbol{v}) \geq \alpha\|\boldsymbol{v}\|_V^2$.

It is readily shown that a solution to the classical formulation (1.3) with the specified boundary condition is also a solution to the weak problem (1.4). Conversely, a solution to the weak problem solves the classical problem provided that the weak solution is sufficiently smooth.

Many problems in mechanics and other areas of physics take the form of variational *inequalities*. These arise in situations, for example, in which a problem is posed on a subset that is not a subspace; or when the model is described by functions that are not differentiable. A classical example of the former is the obstacle problem, in which the deformed shape is sought of a membrane subjected to a transverse force f and which lies above an obstacle described by a continuous function $g : \mathbb{R}^2 \to \mathbb{R}$. The classical formulation of the problem takes the form of a set of complementarity conditions

$$u - g \geq 0\,, \qquad -\Delta u - f \geq 0\,, \qquad (u - g)(\Delta u - f) = 0 \qquad \text{in } \Omega\,. \tag{1.7}$$

Here Δ is the Laplacian operator. These state respectively that the membrane lies on or above the obstacle, the net force on the membrane is nonnegative, and thirdly, that the net force and relative displacement are not simultaneously positive. Assuming once again a homogeneous Dirichlet boundary condition and defining the closed convex set $K = \{\boldsymbol{v} \in H_0^1(\Omega) \mid \boldsymbol{v} \geq g \text{ a.e in } \Omega\}$, the weak formulation of this problem takes the form of an elliptic variational inequality (EVI): find $u \in K$ that satisfies

$$\int_\Omega \nabla u \cdot \nabla(v - u)\, dx \geq \int_\Omega f(v - u)\, dx \qquad \forall v \in K\,. \tag{1.8}$$

A second example of a variational inequality in one that arises as a result of the presence of a nondifferentiable function in its description. Slow steady flows of Bingham fluids

provide an example of such a model. Denote by $\mathbb{M}$ the set of $d \times d$ symmetric matrices or second-rank tensors: that is, $\mathbb{M} = \{\boldsymbol{\sigma} = (\sigma_{ij}) \,|\sigma_{ji} = \sigma_{ij},\ i,j = 1,\ldots,d\}$. Also, define the deviator τ^D of $\boldsymbol{\tau} \in \mathbb{M}$ by $\boldsymbol{\tau}^D = \boldsymbol{\sigma} - (1/d)(\operatorname{tr}\boldsymbol{\tau})\boldsymbol{I}$. Bingham fluids are rigid-viscous fluids for which flow takes place only if the stress deviator $\boldsymbol{\sigma}^D$ exceeds a threshold given by a specified function g. With $\boldsymbol{\varepsilon}$ defined as in (1.6), the flow condition is given in terms of the velocity $\boldsymbol{v}$ by

$$\boldsymbol{\varepsilon}(\boldsymbol{v}) = \begin{cases} \dfrac{1}{2\mu}\left(1 - \dfrac{g}{|\boldsymbol{\sigma}^D|}\right)\boldsymbol{\sigma}^D & \text{if } |\boldsymbol{\sigma}^D| > g\,, \\ 0 & \text{if } |\boldsymbol{\sigma}^D| \le g\,. \end{cases} \tag{1.9}$$

Here μ is the viscosity of the fluid and g is the yield limit. The condition of incompressibility $\operatorname{tr}\boldsymbol{\varepsilon}(\boldsymbol{v}) = \operatorname{div}\boldsymbol{v} = \boldsymbol{0}$ is built into the model by specifying the flow in terms of the stress deviator. The equation for momentum balance

$$-\operatorname{div}\boldsymbol{\sigma} = \boldsymbol{f} \tag{1.10}$$

completes the description of the problem. The corresponding weak formulation of the problem is then as follows: assuming a homogeneous Dirichlet boundary condition and with $V = [H_0^1(\Omega)]^d$ as before, find $\boldsymbol{v} \in V$ that satisfies

$$a(\boldsymbol{v}, \boldsymbol{w} - \boldsymbol{v}) + j(\boldsymbol{w}) - j(\boldsymbol{v}) \ge \langle \ell, \boldsymbol{w} - \boldsymbol{v}\rangle \quad \forall\, v \in V\,. \tag{1.11}$$

Here

$$\begin{aligned} &a : V \times V \to \mathbb{R}, \qquad a(\boldsymbol{v}, \boldsymbol{w}) = \int_\Omega \mu\boldsymbol{\varepsilon}(\boldsymbol{v}) : \boldsymbol{\varepsilon}(\boldsymbol{w})\, dx\,, \\ &j : V \to \mathbb{R}, \qquad\quad j(\boldsymbol{w}) = \int_\Omega g|\boldsymbol{\varepsilon}(\boldsymbol{w})|\, dx\,, \end{aligned} \tag{1.12}$$

and ℓ is as in (1.5b). The Bingham flow problem has been studied mathematically in [5, 6, 21]. Conditions for existence and uniqueness of the solution are given in [5] (Chapter 1, Section 5).

Variational inequalities of the types (1.8) and (1.11) may be formulated in a unified way as follows. Let V be a real Hilbert space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$. Let K be a set in the space V and let $j : K \to \mathbb{R}$. We extend j to all of V by defining

$$j(v) = +\infty \text{ if } v \in V/K. \tag{1.13}$$

Recall that $j : V \to \overline{\mathbb{R}} \equiv \mathbb{R} \cup \{\pm\infty\}$ is proper if $j(v) > -\infty$ for all $v \in V$ and $j(v) \not\equiv \infty$. This property of j is valid if K is nonempty. Also recall that $j : V \to \overline{\mathbb{R}}$ is lower semicontinuous (l.s.c.) if

$$v_n \underset{n\to\infty}{\longrightarrow} v \text{ in } V \quad \Longrightarrow \quad j(v) \le \liminf_{n\to\infty} j(v_n).$$

The extension $j : V \to \overline{\mathbb{R}}$ is l.s.c. if and only if $K \subset V$ is closed and $j : K \to \mathbb{R}$ is l.s.c. Then the general problem becomes one of finding $u \in K$ such that

$$a\,(u, v - u) + j(v) - j(u) \ge \langle \ell, v - u\rangle \quad \forall\, v \in K\,. \tag{1.14}$$

The case (1.8) is recovered by defining $j(v) = 0$ for $v \in K$, while a VI of the kind (1.11) is recovered by setting $K = V$.

We have the following result.

Theorem 1.1. *Let V be a real Hilbert space, $K \subset V$ a non-empty, closed and convex subset, and $a : V \times V \to \mathbb{R}$ a continuous bilinear form with the property that a is $V-$elliptic: that is, $a(v,v) \geq \alpha \|v\|^2$ for some positive constant α. Assume also that $j : K \to \mathbb{R}$ is convex and l.s.c. Then for any $\ell \in V'$, the elliptic variational inequality* (1.14) *has a unique solution. Moreover, the solution u depends Lipschitz continuously on ℓ.*

Returning to problems of evolution, to formulate these properly we need to define spaces of functions as maps from a time interval to a Banach space. Thus, given a Banach space X, $L^p(0,T;X)$ denotes the space of (equivalence classes of) measurable functions from $[0,T]$ to X for which

$$\|f\|_{L^p(0,T;X)} := \Big[\int_0^T \|f\|_X^p \, dx \Big]^{1/p} < \infty \,. \tag{1.15}$$

This is a Banach space with norm defined by (1.15). For integer $m \geq 0$ and real $p \geq 1$, we denote by $W^{m,p}(0,T;X)$ the space of functions $f \in L^p(0,T;X)$ such that the generalized ith time derivative $f^{(i)}$ satisfies $f^{(i)} \in L^p(0,T;X)$. This is a Banach space with the norm

$$\|f\|_{W^{m,p}(0,T;X)} := \Big[\sum_{i=0}^{m} \|f^{(i)}\|_{L^p(0,T;X)} \Big]^{1/p} . \tag{1.16}$$

For the case $m = 0$ we use the conventional notation $W^{0,p}(0,T;X) \equiv L^p(0,T;X)$, while we set $W^{m,2}(0,T;X) \equiv H^m(0,T;X)$.

An example of a *parabolic variational inequality* is the problem of finding

$$u \in L^2(0,T;V) \text{ with } \dot{u} \in L^2(0,T;V') \text{ and } u(0) = u_0,$$

such that for almost all $t \in [0,T]$, $u(t) \in K$ and

$$(\dot{u}(t), v - u(t)) + a(u(t), v - u(t)) \geq \langle f(t), v - u(t) \rangle \quad \forall\, v \in K. \tag{1.17}$$

Conditions for the existence and uniqueness of a solution $u, \dot{u} \in L^2(0,T;V) \cap L^\infty(0,T;H)$ are given in [8, Chapter 6, Section 2], for $f, \dot{f} \in L^2(0,T;V')$ and for some time interval $[0,T]$.

Unsteady slow flows of Bingham fluids provide an example of a parabolic VI. For such a situation the problem (1.11) is generalized to one of finding $\boldsymbol{v} \in L^2(0,T;V)$ such that

$$(\dot{\boldsymbol{v}}(t), \boldsymbol{w} - \boldsymbol{v}(t)) + a(\boldsymbol{v}(t), \boldsymbol{w} - \boldsymbol{v}) + j(\boldsymbol{w}) - j(\boldsymbol{v}) \geq \langle \ell(t), \boldsymbol{w} - \boldsymbol{v}(t) \rangle \quad \forall \boldsymbol{w} \in L^2(0,T;V) \tag{1.18}$$

where $\langle \ell(t), \boldsymbol{v} \rangle = \int_\Omega \boldsymbol{f}(t) \cdot \boldsymbol{v} \, dx$.

Basic results on variational inequalities, including those presented in this section, may be found in [5, 7, 12, 13], for example.

An abstract VI motivated by elastoplasticity. The focus of this work will be on a class of variational inequalities that arise in the context of elastoplasticity, which describes materials whose behaviour is a combination of elasticity and non-reversible path-dependence. The abstract inequality, which is related to but is nontrivially distinct from parabolic VIs such as (1.17), takes the following form: given a Hilbert space W, find $w : [0,T] \to W$, $w(0) = 0$, such that for almost all $t \in (0,T)$, $\dot{w}(t) \in W$ and

$$a(w(t), z - \dot{w}(t)) + j(z) - j(\dot{w}(t)) \geq \langle \ell(t), z - \dot{w}(t) \rangle \quad \forall\, z \in W \,. \tag{1.19}$$

Here $a(\cdot,\cdot)$ and $\ell(\cdot)$ are respectively a bilinear form and linear functional, and $j(\cdot)$ is a positively homogeneous functional. The inequality (1.19) is in fact the differential inclusion

$$Aw(t) - \ell(t) \in \partial j(\dot{w}(t)), \tag{1.20}$$

in which ∂j denotes the subdifferential of j and the operator $A : V \to V'$ is defined by $\langle Aw, z\rangle = a(w, z)$.

It is assumed here that the formulation (1.19) possibly incorporates a situation in which the VI is required to be satisfied on a convex subset $K \subset W$, as for example in (1.17). For such a situation j would be extended from K to all of W as in (1.13), without a change in notation.

Elastoplasticity. We describe the relationship between (1.19) and the problem of elastoplasticity. The variables of interest are the displacement $\boldsymbol{u}$, plastic strain $\boldsymbol{p}$ and a scalar-valued hardening variable η. The problem is described by the equilibrium equation, an elastic relation between the stress $\boldsymbol{\sigma}$ and elastic strain, and a flow relation. The equilibrium equation is

$$-\operatorname{div}\boldsymbol{\sigma} = \boldsymbol{f} \tag{1.21}$$

and the elastic relation is given by

$$\boldsymbol{\sigma}(\boldsymbol{u},\boldsymbol{p}) = \mathsf{C}[\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{p}] := \lambda \operatorname{tr}(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{p}) + 2\mu(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{p}) \tag{1.22}$$

where the total strain $\boldsymbol{\varepsilon}$ is defined in (1.6), $\boldsymbol{p}$ is the plastic strain tensor, and C is the elasticity tensor, given here for isotropic bodies in terms of the strictly positive scalar Lamé parameters λ and μ which were earlier introduced in (1.2).

To describe plastic behaviour we require first the notion of an elastic region: this is a convex region $\mathcal{E} \subset \mathbb{M} \times \mathbb{R}$ given by

$$\mathcal{E} = \{(\boldsymbol{\sigma}, g) \in \mathbb{M} \times \mathbb{R} \mid \varphi(\boldsymbol{\sigma}) + g - c_0 \le 0\}\,. \tag{1.23}$$

The function g is defined as a function of the nonnegative hardening variable η with $g(0) = 0$: for convenience in what follows we assume a linear relationship, so that

$$g(\eta) = -k_2\eta \tag{1.24}$$

in which $k_2 > 0$ is a specified material coefficient.

In its most basic form flow takes place in the direction to the normal to the surface $\mathcal{E}$ when the pair $(\boldsymbol{\sigma}, g)$ lies on the surface $\{(\boldsymbol{\sigma}, g) \in \mathbb{M}\times\mathbb{R} \mid \varphi(\boldsymbol{\sigma}) + g = 0\}$. More compactly,

$$(\dot{\boldsymbol{p}}, \dot{\eta}) \in \mathcal{N}_{\mathcal{E}}(\boldsymbol{\sigma}, g)\,, \tag{1.25}$$

in which the normal cone $\mathcal{N}_{\mathcal{E}}(\boldsymbol{\sigma}, g)$ is defined by

$$\mathcal{N}_{\mathcal{E}}(\boldsymbol{\sigma}, g) = \{(\boldsymbol{q}, \zeta) \mid (\boldsymbol{\tau} - \boldsymbol{\sigma}) : \boldsymbol{q} + (h - g)\zeta \le 0 \ \ \forall(\boldsymbol{\tau}, h) \in \mathcal{E}\}\,. \tag{1.26}$$

A more general form of the flow relation makes provision for translation of the stress in $\mathcal{E}$ by a multiple of the plastic strain. This extension, known as linear kinematic hardening, leads to (1.25) being modified to read

$$(\dot{\boldsymbol{p}}, \dot{\eta}) \in \mathcal{N}_{\mathcal{E}}(\boldsymbol{\sigma} - k_1\boldsymbol{p}, g)\,, \tag{1.27}$$

in which k_1 is a nonnegative scalar. The form (1.27) will be adopted in what follows.

The relation (1.27) may be expressed in an alternative form by introducing the indicator function $I_{\mathcal{E}}$:

$$I_{\mathcal{E}} : \mathbb{M} \to \overline{\mathbb{R}}\,, \qquad I_{\mathcal{E}}(\boldsymbol{\tau}) = \begin{cases} 0 & \boldsymbol{\tau} \in \mathcal{E} \\ +\infty & \text{otherwise}\,. \end{cases} \tag{1.28}$$

Noting that $\mathcal{N}_{\mathcal{E}}$ is equivalent to the subdifferential of the indicator function of $\mathcal{E}$, that is,

$$\mathcal{N}_{\mathcal{E}} = \partial I_{\mathcal{E}}, \tag{1.29}$$

it follows that the relation (1.27) may be inverted in the sense that

$$(\dot{\boldsymbol{p}}, \dot{\eta}) \in \mathcal{N}_{\mathcal{E}}(\boldsymbol{\sigma} - k_1\boldsymbol{p}, g) \Longleftrightarrow (\boldsymbol{\sigma} - k_1\boldsymbol{p}, g) \in \partial I^*_{\mathcal{E}}(\dot{\boldsymbol{p}}, \dot{\eta})\,. \tag{1.30}$$

Here the Legendre-Fenchel conjugate $f^* : X \to \overline{\mathbb{R}}$ of a proper, convex, lsc function f on a normed space X is defined by

$$f^*(x^*) = \sup_{x \in X} \langle x^*, x\rangle - f(x)\,. \tag{1.31}$$

In the language of convex analysis $I^*_{\mathcal{E}}$ is called the support function of $I_{\mathcal{E}}$, while in the context of plasticity theory it is known as the dissipation function, as it characterizes the dissipation or rate of irreversible plastic work.

We take as a simple but physically important example of (1.23) the Mises-Hill condition. For this case, $\varphi(\boldsymbol{\sigma}) = |\boldsymbol{\sigma}^D|$ where, for any $\boldsymbol{\tau} \in \mathbb{M}$, $|\boldsymbol{\tau}|^2 = \sum_{i,j} \tau_{ij}\tau_{ij}$, and as before $\boldsymbol{\sigma}^D := \boldsymbol{\sigma} - (1/d)(\operatorname{tr}\boldsymbol{\sigma})\boldsymbol{I}$ is the deviator of $\boldsymbol{\sigma}$. Then the support or dissipation function is given by

$$I^*_{\mathcal{E}}(\boldsymbol{q}, \zeta) = \begin{cases} c_0|\boldsymbol{q}| & |\boldsymbol{q}| \le \zeta\,, \\ +\infty & \text{otherwise}\,. \end{cases} \tag{1.32}$$

The weak form of the problem of elastoplasticity then follows from (1.21) together with (1.22) to give a weak form of the equilibrium equation; and by expanding (1.30) and integrating over the domain to obtain a weak form for the flow inequality. These two steps lead to the following problem: find $\boldsymbol{u}(t)$ and $\boldsymbol{p}(t)$ that satisfy

$$\int_\Omega \boldsymbol{\sigma}(\boldsymbol{u}, \boldsymbol{p}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx\,, \tag{1.33a}$$

$$\int_\Omega I^*_{\mathcal{E}}(\boldsymbol{q})\, dx \ge \int_\Omega I^*_{\mathcal{E}}(\dot{\boldsymbol{p}})\, dx + \int_\Omega [\boldsymbol{\sigma}(\boldsymbol{u}, \boldsymbol{p}) - k_1\boldsymbol{p} : \boldsymbol{q}] : (\boldsymbol{q} - \dot{\boldsymbol{p}})\, dx + \int_\Omega g(\eta)(\zeta - \dot{\eta})\, dx \tag{1.33b}$$

for all $\boldsymbol{v}$, $\boldsymbol{q}$ and ζ defined in suitable spaces.

The spaces V, Q and M of displacements, plastic strains and hardening variables are defined respectively by

$$\begin{aligned} &V := [H^1_0(\Omega)]^3, \\ &Q := \left\{\boldsymbol{q} = (q_{ij})_{3\times 3} :\ q_{ji} = q_{ij},\ q_{ij} \in L^2(\Omega),\ \operatorname{tr}\boldsymbol{q} = 0 \text{ a.e. in } \Omega\right\}, \\ &M := L^2(\Omega). \end{aligned} \tag{1.34}$$

In the case of Q the inner product is generated by $[L^2(\Omega)]^{3\times 3}$. We set $W := V \times Q \times M$, which is a Hilbert space with the inner product

$$(\mathsf{w}, \mathsf{z})_W := (\boldsymbol{u}, \boldsymbol{v})_V + (\boldsymbol{p}, \boldsymbol{q})_Q + (\eta, \zeta)_M$$

and the norm $\|\mathsf{z}\|_W := (\mathsf{z},\mathsf{z})_W^{1/2}$, where $\mathsf{w} = (\boldsymbol{u},\boldsymbol{p},\eta)$ and $\mathsf{z} = (\boldsymbol{v},\boldsymbol{q},\zeta)$, and define the subset

$$\begin{aligned} W_p &:= \{\mathsf{z} = (\boldsymbol{v},\boldsymbol{q},\zeta) \in W :\ I_{\mathcal{E}}^*(\boldsymbol{q},\zeta) < \infty \text{ a.e. in } \Omega\} \\ &= \{\mathsf{w} \in W :\ |\boldsymbol{q}| \le \zeta \text{ a.e. in } \Omega\} \end{aligned} \tag{1.35}$$

which is a nonempty, closed, convex cone in W.

The problem (1.33) may be cast in the form of the VI (1.19) by setting $\mathsf{w} = (\boldsymbol{u},\boldsymbol{p},\eta)$, $\mathsf{z} = (\boldsymbol{v},\boldsymbol{q},\zeta)$, and defining

$$a : W \times W \to \mathbb{R}, \quad a(\mathsf{w},\mathsf{z}) = \int_\Omega \Big[\boldsymbol{\sigma}(\boldsymbol{u},\boldsymbol{p}) : \boldsymbol{\varepsilon}(\boldsymbol{v}-\boldsymbol{q}) + k_1\boldsymbol{p} : \boldsymbol{q} + k_2\eta\zeta\Big]\, dx\,, \tag{1.36a}$$

$$j : W \times \mathbb{R}, \qquad j(\mathsf{z}) = \int_\Omega I_{\mathcal{E}}^*(\boldsymbol{q})\, dx, \tag{1.36b}$$

$$\mathsf{l} : W \to \mathbb{R}, \qquad \langle \mathsf{l},\mathsf{z}\rangle = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\, dx. \tag{1.36c}$$

Then the problem (1.33) is as follows: find $\mathsf{w} \in W_p$ that satisfies

$$a(\mathsf{w}(t),\mathsf{z}-\dot{\mathsf{w}}(t)) + j(\mathsf{z}) - j(\dot{\mathsf{w}}(t)) \ge \langle \mathsf{l}(t),\mathsf{z}-\dot{\mathsf{w}}(t)\rangle \quad \forall\, \mathsf{z} \in W_p\,. \tag{1.37}$$

In the following section we will review the conditions under which the abstract problem (1.19) has a unique solution. Thereafter the focus of the work will be on issues pertaining to approximations of the VI. The first topic in this context will be that of convergence of fully discrete approximations based on the use of finite elements in space. Thereafter, attention will shift to the construction and analysis of an algorithm for determining approximate solutions. The essence of the algorithm is a predictor-corrector approach which in the context of elastoplasticity is suggested by the pair of relations characterizing the problem: an equation of equilibrium, and a flow relation that takes the form of an inequality. We will also discuss the interpretation of the algorithm, notably the corrector step which is referred to in the computational plasticity literature as a return map.

2. Well-posedness of the abstract VI

We return to (1.19) and set out conditions for existence of a unique solution. The bilinear form $a : W \times W \to \mathbb{R}$ is assumed to symmetric, bounded, and W-elliptic: that is,

$$a(w,z) = a(z,w) \quad \forall\, w,z \in W,$$

and there exist constants $c_0, c_1 > 0$ such that

$$|a(w,z)| \le c_1\|w\|_W\|z\|_W, \ \ a(z,z) \ge c_0\|z\|_W^2 \quad \forall\, w,z \in W.$$

Theorem 2.1. *Let W be a Hilbert space; $K \subset W$ a nonempty, closed, convex cone; a: $W\times W \to \mathbb{R}$ a bilinear form that is symmetric, bounded, and W-elliptic; $\ell \in H^1(0,T;W')$ with $\ell(0) = 0$; and $j : K \to \mathbb{R}$ nonnegative, convex, positively homogeneous, and Lipschitz continuous. Then there exists a unique solution w of* (1.19) *satisfying $w \in H^1(0,T;W)$.*

Furthermore, $w : [0,T] \to W$ *is the solution to problem* (1.19) *if and only if there is a function* $w^*(t)$*:* $[0,T] \to W'$ *such that for almost all* $t \in (0,T)$*,*

$$a(w(t),z) + \langle w^*(t), z\rangle = \langle \ell(t), z\rangle \quad \forall\, z \in W, \tag{2.1}$$

$$w^*(t) \in \partial j(\dot{w}(t)). \tag{2.2}$$

Remark 2.2. Questions of existence and uniqueness of solutions to this problem were first presented in [22] in the context of elastoplasticity for the case $k_2 = 0$. The results were extended in [11] to the more general problem.

We observe from (2.1) that w^* has the regularity property

$$w^* \in H^1(0,T;W'). \tag{2.3}$$

The proof of existence involves two stages: the first entails discretizing in time and establishing the existence of a family of solutions $\{w_n\}_{n=1}^N$ to the discrete problem. Time-discretization involves a uniform partitioning of the time interval $[0,T]$ according to

$$0 = t_0 < t_1 < \cdots < t_N = T, \ \text{ where } \ t_n - t_{n-1} = k, \ \ k = T/N.$$

We write $\ell_n = \ell(t_n)$. Corresponding to a sequence $\{w_n\}_{n=0}^N$, we define Δw_n to be the backward difference $w_n - w_{n-1}$, and $\delta w_n = \Delta w_n/k$ to be the backward divided difference, $n = 1,2,\ldots,N$. The semidiscrete counterpart of (1.19) is then as follows: given $\{\ell_n\}_{n=0}^N \subset H'$, $\ell_0 = 0$, find $\{w_n\}_{n=0}^N \subset W$ with $w_0 = 0$ that satisfies $\Delta w_n \in K$ and

$$a(w_n, z - \Delta w_n) + j(z) - j(\Delta w_n) \geq \langle \ell_n, z - \Delta w_n\rangle \quad \forall\, z \in W\,, \qquad n = 1,2,\ldots,N. \tag{2.4}$$

The second stage involves the construction of piecewise linear interpolants w^k of the discrete solutions $\{w_n\}_{n=1}^N$ and showing that the limit of these interpolants as the time step-size k approaches zero is in fact a solution of (1.19).

Remark 2.3. Problem (1.37) is readily shown to satisfy the conditions of Theorem 2.1 and to have a unique solution provided that $\bar{k}_i := \operatorname{essinf}_\Omega k_i > 0$ for $i = 1$ or $i = 2$.

3. Fully discrete approximations of the abstract problem

We present in this section an overview of results on the convergence of fully discrete approximations of the problem (1.19). Time discretization is carried out as before to arrive at the semidiscrete problem (2.4). In addition, we define a family of finite-dimensional subspaces $W^h \subset W$ parametrized by $h > 0$, with the property that

$$\lim_{h\to 0} \inf_{z^h \in W^h} \|z - z^h\|_W = 0 \quad \forall\, z \in W. \tag{3.1}$$

Set $K^h = W^h \cap K$, which is nonempty, since $0 \in K^h$. Furthermore, K^h is a nonempty, closed, convex cone in W^h, and in W as well.

Let $\theta \in [\frac{1}{2}, 1]$ be a parameter. The fully discrete approximation problem is as follows: find $w^{hk} = \{w_n^{hk}\}_{n=0}^N$, where $w_n^{hk} \in W^h$, $0 \leq n \leq N$, with $w_0^{hk} = 0$, such that for $n = 1,2,\ldots,N$, $\delta w_n^{hk} \in K^h$ and for all $z^h \in K^h$,

$$a(\theta\, w_n^{hk} + (1-\theta)\, w_{n-1}^{hk}, z^h - \delta w_n^{hk}) + j(z^h) - j(\delta w_n^{hk}) \geq \langle \ell_{n-1+\theta}, z^h - \delta w_n^{hk}\rangle\,. \tag{3.2}$$

This problem admits a unique solution $w^{hk} \subset K^h$.

The reason for restricting attention to the interval $\theta \in [\frac{1}{2}, 1]$ is as follows. The case $\theta = 1$ corresponds to a backward Euler approximation of first derivatives in time, while for $\theta = \frac{1}{2}$ we have the Crank-Nicolson scheme. The choice $\theta > 1$ is not good, for one would have to use a value of ℓ outside the time interval $[0, T]$. The case $\theta = 0$ is degenerate. The choice $0 \neq \theta < \frac{1}{2}$ leads to an unstable scheme, as is easily seen by setting $j = 0$ and $K = W$. This choice yields a linear problem, and an analysis along the lines of that in [24], for example, shows that a perturbation ϵ in the initial value leads to a perturbation error e with magnitude

$$|e| = \Big(\frac{1-\theta}{\theta}\Big)^n \epsilon;$$

for $\theta < \frac{1}{2}$, $|e| \to \infty$ as $n \to \infty$ and so small perturbations in the initial conditions result in arbitrarily large errors in the approximation.

The quantity of interest is the error $e_n = w_n - w_n^{hk}$, $0 \leq n \leq N$. The following error estimate is obtained in [10], §11.3.

Theorem 3.1. *Suppose that W, K, a, ℓ, and j satisfy the assumptions in Theorem 2.1. Let $w \in H^1(0,T;W)$ and w^{hk} be the solutions to (1.19) and (3.2) respectively. Then if $w \in W^{3,1}(0,T;W)$, the estimates*

$$\max_n \|w_n - w_n^{hk}\|_W \leq c\,k + c\,E_\theta^{hk} \quad \text{if } \theta \neq \frac{1}{2} \tag{3.3}$$

and

$$\max_n \|w_n - w_n^{hk}\|_W \leq c\,k^2 + c\,E_{1/2}^{hk} \quad \text{if } \theta = \frac{1}{2} \tag{3.4}$$

hold, where

$$E_\theta^{hk} = \inf_{\substack{z_j^h \in K^h \\ j=1,\ldots,N}} \left\{ k \sum_{j=1}^N \|\dot{w}_{j-1+\theta} - z_j^h\|_W + \Big[k \sum_{j=1}^N R(\dot{w}_{j-1+\theta}, z_j^h) \Big]^{1/2} \right\} \tag{3.5}$$

and

$$R(\dot{w}_{j-1+\theta}, z_j^h) = a(w_{n-1+\theta}, \dot{w}_{j-1+\theta} - z_j^h) + j(\dot{w}_{j-1+\theta}) - j(z_j^h) - \langle \ell_{n-1+\theta}, \dot{w}_{j-1+\theta} - z_j^h \rangle\,. \tag{3.6}$$

Remark 3.2. The orders are optimal with respect to the time step-size in the error estimates (3.3) and (3.4). In particular, for the backward Euler scheme with $\theta = 1$ the approximation is of first order in time, while the Crank–Nicolson-type scheme with $\theta = \frac{1}{2}$ gives second-order accuracy.

Elastoplasticity. We return to the problem (1.37). In the context of finite element approximations the space W^h is defined by first constructing a partition or mesh $\mathcal{T}$ of Ω, assumed for convenience to be polygonal or polyhedral, into triangles (resp. tetrahedra) such that $\overline{\Omega} = \cup_{\mathcal{T}} T$. Any two distinct members T_1 and T_2 of $\mathcal{T}$ are either disjoint or share a common vertex, edge or, in the case $d = 3$, a common face. Set $h_T = \max\{|\boldsymbol{x} - \boldsymbol{y}|\,,\ \boldsymbol{x}\,, \boldsymbol{y} \in T\}$ and denote by ρ_T the diameter of the largest disc (for $d = 2$) or sphere (for $d = 3$) contained in T. The mesh $\mathcal{T}$ is assumed to be shape-regular in the sense that there exists a constant β such that $h_T/\rho_T \leq \beta$ for all $T \in \mathcal{T}$. We define the mesh size $h = \max_{T \in \mathcal{T}} h_T$. Set

$$P_k(T) := \{v : T \to \mathbb{R}^d \mid v \text{ is a polynomial of degree} \leq k \text{ on } T\}\,. \tag{3.7}$$

Let $W^h := V^h \times Q^h \times M^h$ be a finite-dimensional subspace of W, and set $K^h := W^h \cap K = V^h \times K_0^h$, where

$$K_0^h := \{(\boldsymbol{q}^h, \zeta^h) \in Q^h \times M^h : |\boldsymbol{q}^h| \leq \zeta^h \text{ in } \Omega\}.$$

We choose

$$V^h = \{\boldsymbol{v} = (v_i) \in V \mid v_i \in C(\overline{\Omega}) \mid v_i|_T \in P_1(T)\}\,, \tag{3.8a}$$

$$Q_h = \{\boldsymbol{q} = (q_{ij}) \in Q \mid (q_{ij})|_T \in P_0(T)\}\,, \tag{3.8b}$$

$$M_h = \{\zeta \in M \mid \zeta|_T \in P_0(T)\}\,. \tag{3.8c}$$

Assume that $\dot{\boldsymbol{u}} \in L^2(0,T;[H^2(\Omega)]^d)$. Then from the standard interpolation error estimates for finite elements (see for example [2]), we have

$$\inf_{\boldsymbol{v}^h \in L^2(0,T;V^h)} \|\dot{\boldsymbol{u}} - \boldsymbol{v}^h\|_{L^2(0,T;V)} \leq c\,h. \tag{3.9}$$

Further, assume that $\dot{\boldsymbol{p}} \in L^2(0,T;[H^1(\Omega)]^{d\times d})$, and $\dot{\eta} \in L^2(0,T;H^1(\Omega))$. Let $\boldsymbol{q}^h = \Pi^h \dot{\boldsymbol{p}}$ be the orthogonal projection of $\dot{\boldsymbol{p}}$ onto Q^h with respect to the inner product of Q. Similarly, we take $\zeta^h = \Pi^h \dot{\eta}$ to be the orthogonal projection of $\dot{\eta}$ onto M^h with respect to the inner product of M. Since $\dot{\mathsf{w}} \in K$ and K is convex, we have $(\Pi^h \dot{\boldsymbol{p}}, \Pi^h \dot{\eta}) \in K_0^h$, and standard interpolation estimates again give

$$\|\dot{\boldsymbol{p}} - \Pi^h \dot{\boldsymbol{p}}\|_{L^2(0,T;Q)} \leq c\,h, \tag{3.10a}$$

$$\|\dot{\eta} - \Pi^h \dot{\eta}\|_{L^2(0,T;M)} \leq c\,h. \tag{3.10b}$$

In the backward Euler approximation of the problem $\mathsf{w}_0^{hk} = \mathsf{0}$ and we compute $\mathsf{w}_n^{hk} = (\boldsymbol{u}_n^{hk}, \boldsymbol{p}_n^{hk}, \eta_n^{hk}) : [0,T] \to W^h$, $n = 1, 2, \ldots, N$, such that $\delta \mathsf{w}_n^{hk} \in K^h$ and

$$a(\mathsf{w}_n^{hk}, \mathsf{z}^h - \delta \mathsf{w}_n^{hk}) + j(\mathsf{z}^h) - j(\delta \mathsf{w}_n^{hk}) \geq \langle \mathsf{l}_n, \mathsf{z}^h - \delta \mathsf{w}_n^{hk} \rangle \quad \forall\, \mathsf{z}^h \in K^h. \tag{3.11}$$

This problem has a unique solution. The quantity R defined in (3.6) can be shown to reduce to a term involving $j(\mathsf{z}^h) - j(\dot{\mathsf{w}}_{j-1+\theta})$ which depends only on $\dot{\boldsymbol{p}}$ and $\boldsymbol{q}^h$. Thus we find that if $\ddot{\mathsf{w}} \in L^2(0,T;W)$, then

$$\max_{0\leq n\leq N} \|\mathsf{w}_n - \mathsf{w}_n^{hk}\|_W^2 \leq c\,k^2 + c\,k \sum_{n=1}^{N} \Big[\inf_{\boldsymbol{v}^h \in V^h} \|\dot{\boldsymbol{u}}_n - \boldsymbol{v}^h\|_V^2 + \inf_{(\boldsymbol{q}^h, \zeta^h) \in K_0^h} \big(\|\dot{\boldsymbol{p}}_n - \boldsymbol{q}^h\|_Q + \|\dot{\eta}_n - \zeta^h\|_M^2 \big) \Big]. \tag{3.12}$$

The interpolation estimates (3.9) and (3.10) lead to the error bound

$$\max_{0\leq n\leq N} \|\mathsf{w}_n - \mathsf{w}_n^{hk}\|_W \leq c\,(h^{1/2} + k). \tag{3.13}$$

Similarly, for the Crank–Nicolson scheme and suitable smoothness assumptions on the solution of the original problem the error estimate is

$$\max_{0\leq n\leq N} \|\mathsf{w}_n - \mathsf{w}_n^{hk}\|_W \leq c\,(h^{1/2} + k^2)\,. \tag{3.14}$$

Optimal-order estimates. The $O(h^{1/2})$ convergence rate in (3.13) and (3.14) for lowest-order polynomial approximations is determined by the term involving the interpolation of the plastic strain rate $\dot{\boldsymbol{p}}$ on the right-hand side of (3.12). An assumption of higher regularity of the plastic strain and a higher-order approximation would of course improve the estimate to $O(h)$. In an alternative approach in [1], aimed at obtaining $O(h)$ estimates, the material functions such as λ and μ in (1.22) are approximated by their constant average values on each element in the discrete formulation. It is then shown that for the case of piecewise-constant data and assuming exact integration of the integral involving the loading term, convergence at the optimal $O(h)$ rate is obtained.

A comment on convergence under minimal regularity. The above error analysis assumes a certain degree of regularity of the solution to the original problem. Regularity results have been established for problems in elastoplasticity: for example, in [14, 15] the displacement components are shown under certain conditions to belong to $H^{3/2-\delta}(\Omega)$ and the components of plastic strain and hardening variable to $H^{1/2-\delta}(\Omega)$, for small $\delta > 0$. It is nevertheless of interest to show convergence of the various numerical schemes under the minimal regularity condition of the weak solution.

Recall that the problem (1.19) has a unique solution $w \in H^1(0,T;W)$. Given that $C^\infty([0,T],W)$ is dense in $H^1(0,T;W)$, it follows that for any $\varepsilon > 0$ there exists a function $\hat{w} \in C^\infty([0,T];W)$ such that

$$\|w - \hat{w}\|_{C([0,T];W)} \leq c\,\varepsilon. \tag{3.15}$$

By approximating the solution arbitrarily closely with smooth functions and through a judicious use of Taylor expansions and density arguments, it has been shown in [9] (see also [10], §11.4) that the fully discrete solution w_n^{hk} converges to $w \in H^1(0,T;W)$ in the sense that

$$\max_{1\leq n\leq N} \|w_n^{hk} - w_n\|_W \to 0 \quad \text{as } h,k \to 0. \tag{3.16}$$

4. Solution algorithms

We turn next to the question of constructing convergent and efficient solution algorithms. The emphasis here is on the solution of the time-discrete variational inequality (2.4), which in the context of this section could be assumed to be a semidiscrete approximation or the fully discrete version (3.2) with the backward Euler assumption $\theta = 1$. For convenience we will focus on the problem in the form (2.4), which by a rearrangement of terms can be written in the following form: with w_{n-1} known, find $w_n \in K$ such that

$$a(\Delta w_n, z - \Delta w_n) + j(z) - j(\Delta w_n) \geq \langle L_n, z - \Delta w_n\rangle \quad \forall\, z \in K, \tag{4.1}$$

where the functional L_n is defined by

$$\langle L_n, z\rangle := \langle \ell_n, z\rangle - a(w_{n-1}, z)\,. \tag{4.2}$$

The objective is to present and discuss a predictor-corrector approach that has its origins in and exploits the particular structure of the problem of elastoplasticity: see, for example, [17, 25]. In this context we recognise that members of the space W are pairs of the form

$(\boldsymbol{u}, (\boldsymbol{p}, \eta))$ in which $\boldsymbol{u}$ is the displacement and the pair $(\boldsymbol{p}, \eta)$ represents the plastic strain and hardening variable. Thus W is a product space of the form $W := V \times \Lambda$. With this decomposition in mind, for the abstract problem (4.1) we define Hilbert spaces V and Λ and set $W = V \times \Lambda$, $w = (u, \lambda)$, and $z = (v, \zeta)$.

Next, in order to structure the algorithm as one of predictor-corrector type we will decompose the VI (4.1) in a manner corresponding to the structure of the problem (1.37) for elastoplasticity: this problem is written equivalently as the equation (1.33a) and inequality (1.33b). To do likewise with the abstract problem we use the bilinearity of $a(\cdot, \cdot)$ to define continuous bilinear forms $b : V \times V \to \mathbb{R}$, $c : \Lambda \times V \to \mathbb{R}$ and $d : \Lambda \times \Lambda \to \mathbb{R}$, according to

$$a(w, z) = b(u, v) - c(\lambda, v) - c(\zeta, u) + d(\lambda, \zeta)\,. \tag{4.3}$$

We also introduce continuous linear forms, $\ell_1 : V \to \mathbb{R}$ and $\ell_2 : \Lambda \to \mathbb{R}$, and a functional, $j : \Lambda \to \mathbb{R}$, with j assumed to be nonnegative, convex, and Lipschitz continuous, and of the form

$$j(\zeta) := \int_\Omega D(\zeta(\boldsymbol{x}))\, dx\,.$$

The function D is not differentiable at $\zeta = 0$ and is at least twice differentiable everywhere else.

The problem is then as follows: find $u \in V$ and $\lambda \in \Lambda$ such that

$$b(u, v) - c(\lambda, v) = \langle \ell_1, v\rangle \quad \forall\, v \in V, \tag{4.4}$$

$$j(\zeta) - j(\lambda) - c(\zeta - \lambda, u) + d(\lambda, \zeta - \lambda) \geq \langle \ell_2, \zeta - \lambda\rangle \quad \forall\, \zeta \in \Lambda. \tag{4.5}$$

For the problem of elastoplasticity we have $w = (\boldsymbol{u}, (\boldsymbol{p}, \eta))$ so that the space Λ is given by

$$\Lambda := Q \times M \tag{4.6}$$

with Q and M defined in (1.34). The bilinear form $a(\cdot, \cdot)$ and linear functional $L(\cdot)$ corresponding to the incremental problem are found from (4.1) and (4.2), and are

$$a(\mathsf{w}, \mathsf{z}) := \int_\Omega \Big[\mathsf{C}(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{p}) : (\boldsymbol{\varepsilon}(\boldsymbol{v}) - \boldsymbol{q}) + k_1 \boldsymbol{p} : \boldsymbol{q} + k_2 \eta\zeta\Big] dx, \tag{4.7}$$

$$\langle \mathsf{L}_n, \mathsf{z}\rangle := \langle \mathsf{l}_n, \mathsf{z}\rangle - a(\mathsf{w}^k_{n-1}, \mathsf{z})\,. \tag{4.8}$$

It follows that the bilinear forms appearing in the algorithmic formulation are given by

$$b : V \times V \to \mathbb{R}, \quad b(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega \mathsf{C}\,\boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx, \tag{4.9}$$

$$c : Q \times V \to \mathbb{R}, \quad c(\boldsymbol{q}, \boldsymbol{v}) := \int_\Omega \mathsf{C}\,\boldsymbol{q} : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx, \tag{4.10}$$

$$d : (Q \times M) \times (Q \times M) \to \mathbb{R},$$

$$d(\boldsymbol{p}, \eta; \boldsymbol{q}, \zeta) := \int_\Omega (\mathsf{C}\,\boldsymbol{p} : \boldsymbol{q} + k_1 \boldsymbol{p} : \boldsymbol{q} + k_2 \eta\zeta)\, dx. \tag{4.11}$$

The linear functional $\mathsf{L}_n(\cdot)$ may likewise be decomposed by writing it in the form

$$\langle \mathsf{L}_n, \mathsf{z}\rangle := \langle \boldsymbol{L}_{n,1}, \boldsymbol{v}\rangle + \langle \boldsymbol{L}_{n,2}, \boldsymbol{q}\rangle,$$

in which

$$\boldsymbol{L}_{n,1} : V \to \mathbb{R}, \quad \langle \boldsymbol{L}_{n,1}, \boldsymbol{v} \rangle := \int_\Omega [\boldsymbol{f}_n \cdot \boldsymbol{v} - \boldsymbol{\sigma}^k_{n-1} : \boldsymbol{\varepsilon}(\boldsymbol{v})]\, dx$$

and

$$\boldsymbol{L}_{n,2} : Q \to \mathbb{R}, \quad \langle \boldsymbol{L}_{n,2}, \boldsymbol{q} \rangle := \int_\Omega \boldsymbol{\chi}^k_{n-1} : \boldsymbol{q}\, dx,$$

where

$$\begin{aligned} \boldsymbol{\sigma}^k_{n-1} &:= \mathsf{C}\,(\boldsymbol{\varepsilon}(\boldsymbol{u}^k_{n-1}) - \boldsymbol{p}^k_{n-1}), \\ \boldsymbol{\chi}^k_{n-1} &:= \boldsymbol{\sigma}^k_{n-1} + k_1 \boldsymbol{p}^k_{n-1} \end{aligned}$$

are known from the previous step of the computation. □

The solution algorithm. In the predictor–corrector strategy we have estimates u^{i-1} and λ^{i-1}, and we seek new, improved estimates u^i and λ^i.

Here we prove convergence of the algorithm under a general set of conditions: continuity of the bilinear forms $b(\cdot,\cdot)\,, c(\cdot,\cdot)$ and $d(\cdot,\cdot)$ are assumed, as are the continuity of the linear functionals ℓ_1 and ℓ_2, and the conditions on $j(\cdot)$. But the assumption of $(V \times \Lambda)$-ellipticity of $a(\cdot,\cdot)$ is replaced by the weaker requirements of V-ellipticity of $c(\cdot,\cdot)$ and Λ-ellipticity of $d(\cdot,\cdot)$, with no assumptions of symmetry of these bilinear forms. This would allow for situations, for example, in which discrete approximations such as certain discontinuous Galerkin formulations lead to non-symmetric bilinear forms (see for example [3, 4]). An immediate consequence of the lack of symmetry is that there does not exist an equivalent minimization problem, so that the proof of convergence must rely on the formulation (4.4)–(4.5).

We begin by introducing various assumptions.

(i) Let V and Λ be two Hilbert spaces.

(ii) Let $b : V \times V \to \mathbb{R}$, $c : V \times \Lambda \to \mathbb{R}$ and $d : \Lambda \times \Lambda \to \mathbb{R}$ be continuous bilinear forms, with b and d elliptic but not necessarily symmetric. Thus, for some positive constants b_1, b_0, c_1, d_1 and d_0,

$$\begin{aligned} &|b(u,v)| \le b_1 \|u\|_V \|v\|_V, \quad b(u,u) \ge b_0 \|u\|_V^2\,, \\ &|c(u,\lambda)| \le c_1 \|u\|_V \|\lambda\|_\Lambda, \\ &|d(\lambda,\zeta)| \le d_1 \|\lambda\|_\Lambda \|\zeta\|_\Lambda\,, \quad d(\lambda,\lambda) \ge d_0 \|\lambda\|_\Lambda^2 \end{aligned} \tag{4.12}$$

for all $u, v \in V$ and $\lambda, \zeta \in \Lambda$.

(iii) Let $\ell_1 : V \to \mathbb{R}$ and $\ell_2 : \Lambda \to \mathbb{R}$ be continuous linear forms.

(iv) Let $j : \Lambda \to \mathbb{R}$ be a nonnegative, convex and Lipschitz continuous functional.

(v) For $w = (u, \lambda),\ z = (v, \zeta) \in V \times \Lambda$ we define $a(\cdot,\cdot)$ by (4.3).

Note that with these conditions the abstract problem (4.4)–(4.5) has a unique solution $w = (u, \lambda) \in V \times \Lambda$.

The abstract algorithm is given in the following general form:

Given $w^0 = (u^0, \lambda^0) \in V \times \Lambda$, for $i = 1, 2, \ldots,$

Predictor: Compute $(u^i, \lambda^{*i}) \in V \times \Lambda$ such that

$$b(u^i, v) - c(\lambda^{*i}, v) = \langle \ell_1, v \rangle \quad \forall\, v \in V\,, \tag{4.13}$$

$$j^{(i)}(\zeta) - j^{(i)}(\lambda^{*i}) + d(\lambda^{*i}, \zeta - \lambda^{*i}) \\ \geq \langle \ell_2, \zeta - \lambda^{*i} \rangle + c(\zeta - \lambda^{*i}, u^i) \quad \forall\, \zeta \in \Lambda\,. \tag{4.14}$$

Corrector: Compute $\lambda^i \in \Lambda$ such that

$$j(\zeta) - j(\lambda^i) + d(\lambda^i, \zeta - \lambda^i) \geq \langle \ell_2, \zeta - \lambda^i \rangle + c(\zeta - \lambda^i, u^i) \quad \forall\, \zeta \in \Lambda \tag{4.15}$$

where

$$j^{(i)}(\zeta) := \int_\Omega D^{(i)}(\zeta)\, dx$$

with $D^{(i)}$ a smooth convex function, satisfying

$$D^{(i)}(\lambda^{i-1}) = D(\lambda^{i-1})\,, \tag{4.16}$$

$$\nabla D^{(i)}(\lambda^{i-1}) = \nabla D(\lambda^{i-1})\,, \tag{4.17}$$

$$D(\zeta) \leq D^{(i)}(\zeta) \quad \forall\, \zeta \in \Lambda\,. \tag{4.18}$$

Some examples of commonly used predictors follow.

The elastic predictor. For this simple case we take $\lambda^{*i} = \lambda^{i-1}$ and there is no need to define the functional $j^{(i)}$.

While the most straightforward, the use of the elastic predictor leads to slow convergence, so that various alternatives are preferred. Two of these are summarized next.

The secant predictor. The algorithm corresponding to the secant predictor is obtained by choosing $D^{(i)}$ to be the quadratic function whose graph lies inside the cone with boundary the graph of D, and satisfying (4.16)–(4.17).

More precisely, for Λ a space of n-tuples we seek a vector a and a symmetric positive definite matrix B such that the function

$$D^i(\zeta) = D(\lambda^{i-1}) + \mathsf{a} \cdot (\zeta - \lambda^{i-1}) + \tfrac{1}{2}(\zeta - \lambda^i) : \mathsf{B}(\zeta - \lambda^{i-1})$$

satisfies (4.16)–(4.18). We find that $\mathsf{a} = \nabla D(\lambda^{i-1})$ and B is to be chosen such that

$$D(\zeta) \leq D(\lambda^{i-1}) + \nabla D(\lambda^{i-1})(\zeta - \lambda^{i-1}) + \tfrac{1}{2}(\zeta - \lambda^i) : \mathsf{B}(\zeta - \lambda^{i-1})$$

at least in a small neighbourhood of λ^{i-1}. Then all of the conditions (4.16)–(4.18) are satisfied.

The consistent tangent predictor. This predictor is constructed by considering the following modified second order Taylor expansion of the function D about λ^{i-1}:

$$D^i(\zeta) = D(\lambda^{i-1}) + \nabla D(\lambda^{i-1}) \cdot (\zeta - \lambda^{i-1}) \\ + \tfrac{1}{2}(\zeta - \lambda^{i-1}) : [\mathsf{H}(\lambda^{i-1}) + \varepsilon \mathsf{I}](\zeta - \lambda^{i-1})\,. \tag{4.19}$$

Here, $\mathsf{H}(\lambda^{i-1})$ is the Hessian matrix of D at λ^{i-1} and I is the identity matrix. In order that (4.18) be satisfied it is essential that $\varepsilon > 0$, the magnitude of ε being chosen so that D^i satisfies (4.18) at least in a small neighbourhood of λ^{i-1}.

The definition (4.19), without the perturbation, leads in the spatially discrete case to the consistent tangent predictor [26] favoured in computational approaches. As the convergence analysis in the next section will show, the perturbation is necessary to guarantee convergence of the algorithm.

4.1. Convergence analysis of the solution algorithms. We return to the general problem (4.4)–(4.5) and establish conditions for its convergence. This result was first given in [4]; see also [10].

Theorem 4.1. *Under the assumptions on the bilinear forms, functionals and the structural inequality*

$$r_1 := \frac{c_1^2}{b_0 d_0} < \frac{1}{3} \tag{4.20}$$

where c_1, b_0 and d_0 are defined in (4.12), the predictor-corrector algorithm (4.13)–(4.15) converges. That is,

$$u^i \to u \text{ in } V \quad \text{and} \quad \lambda^i \to \lambda \text{ in } \Lambda \quad \text{as } i \to \infty\,,$$

where $w = (u, \lambda)$ is the solution of the abstract problem (4.4)–(4.5)*. Furthermore, the error estimate*

$$\|w^i - w\|_{V\times\Lambda} \le r_0 \frac{2r_1}{1-r_1} \|u^i - u^{i-1}\|_V \tag{4.21}$$

holds, where $w^i = (u^i, \lambda^i)$ and r_0 is defined by

$$r_0 := \left(1 + \frac{c_1^2}{d_0^2}\right)^{1/2}. \tag{4.22}$$

Proof. A sketch of the proof follows.

First, by using the coercivity of b and the continuity of c it can be shown that

$$b_0 \|u^i - u^{i-1}\|_V \le c_1 \|\lambda^{*i} - \lambda^{*(i-1)}\|_\Lambda,$$

or

$$\|u^i - u^{i-1}\|_V \le \frac{c_1}{b_0} \left[\|\lambda^{i-1} - \lambda^{i-2}\|_\Lambda + \|\lambda^{*i} - \lambda^{i-1}\|_\Lambda + \|\lambda^{*(i-1)} - \lambda^{i-2}\|_\Lambda\right]. \tag{4.23}$$

Next, from the properties (4.16) and (4.18) it follows that

$$d(\lambda^{*i} - \lambda^{i-1}, \lambda^{*i} - \lambda^{i-1}) \le c(\lambda^{*i} - \lambda^{i-1}, u^i - u^{i-1}).$$

The coercivity of d and the continuity of c give

$$\|\lambda^{*i} - \lambda^{i-1}\|_\Lambda \le \frac{c_1}{d_0} \|u^i - u^{i-1}\|_V\,. \tag{4.24}$$

Combining (4.23) and (4.24) we get

$$(1 - r_1)\, \|u^i - u^{i-1}\|_V \le \frac{c_1}{b_0} \|\lambda^{i-1} - \lambda^{i-2}\|_\Lambda + r_1 \|u^{i-1} - u^{i-2}\|_V\,, \tag{4.25}$$

where r_1 is defined in (4.20).

Now we take $\zeta = \lambda^{i-1}$ in (4.15), and then replace i by $i-1$ and take $\zeta = \lambda^i$ in (4.15), add the two resulting inequalities and use the coercivity of d and the continuity of c to get

$$\|\lambda^i - \lambda^{i-1}\|_\Lambda \le \frac{c_1}{d_0} \|u^i - u^{i-1}\|_V\,. \tag{4.26}$$

Combining (4.25) and (4.26) we then obtain

$$(1 - r_1)\,\|u^i - u^{i-1}\|_V \leq 2\,r_1 \|u^{i-1} - u^{i-2}\|_V\,,$$

which gives

$$\|u^i - u^{i-1}\|_V \leq \frac{2r_1}{1 - r_1} \|u^{i-1} - u^{i-2}\|_V \tag{4.27}$$

with r_1 defined in (4.20).

An induction procedure based on (4.27) leads to

$$\|u^i - u^{i-1}\|_V \leq \left(\frac{2r_1}{1 - r_1}\right)^{i-1} \|u^1 - u^0\|_V\,. \tag{4.28}$$

Using this bound and (4.26), we have

$$\|\lambda^i - \lambda^{i-1}\|_\Lambda \leq \frac{c_1}{d_0} \left(\frac{2r_1}{1 - r_1}\right)^{i-1} \|u^1 - u^0\|_V\,. \tag{4.29}$$

Therefore,

$$\begin{aligned} \|w^i - w^{i-1}\|_{V\times\Lambda} &= \left(\|u^i - u^{i-1}\|_V^2 + \|\lambda^i - \lambda^{i-1}\|_\Lambda^2\right)^{1/2} \\ &\leq r_0 \left(\frac{2r_1}{1 - r_1}\right)^{i-1} \|u^1 - u^0\|_V\,, \end{aligned} \tag{4.30}$$

where r_0 is defined in (4.22). Since $r_1 < 1/3$, $2r_1/(1 - r_1) < 1$ and thus $\{w^i\}_{i\geq 1}$ is a Cauchy sequence in the Hilbert space $V \times \Lambda$, converging to some limit $w^* = (u^*, \lambda^*) \in V \times \Lambda$.

Using the continuity of the bilinear and linear forms we can pass to the limit in (4.13) and (4.15) and find that $w^* = (u^*, \lambda^*)$ solves the abstract problem (4.4) and (4.5). By the uniqueness of the solution it follows that $w^* = w$. Therefore the sequence $\{w^i\}_{i\geq 1}$ converges to w.

The estimate (4.21) follows by a further lengthy but straightforward process to show that

$$\|u^i - u\|_V \leq r_1 \|u^i - u^{i-1}\|_V + \frac{c_1}{b_0} \|\lambda^{i-1} - \lambda\|_\Lambda. \tag{4.31}$$

We similarly obtain

$$\|\lambda^i - \lambda\|_\Lambda \leq \frac{c_1}{d_0} \|u^i - u\|_V\,. \tag{4.32}$$

Combining (4.31) and (4.32), we have

$$\begin{aligned} \|u^i - u\|_V &\leq r_1 \|u^i - u^{i-1}\|_V + r_1 \|u^{i-1} - u\|_V \\ &\leq 2\,r_1 \|u^i - u^{i-1}\|_V + r_1 \|u^i - u\|_V \end{aligned}$$

and hence,

$$\|u^i - u\|_V \leq \frac{2\,r_1}{1 - r_1} \|u^i - u^{i-1}\|_V\,. \tag{4.33}$$

Then with (4.32), we further have

$$\|\lambda^i - \lambda\|_\Lambda \leq \frac{c_1}{d_0} \frac{2\,r_1}{1 - r_1} \|u^i - u^{i-1}\|_V\,. \tag{4.34}$$

Using (4.33) and (4.34) we obtain (4.21). This completes the proof. □

For the elastoplasticity problem (1.36) with (1.37), the relevant bilinear forms are given by (4.7)–(4.11). The space V of displacements is as before, and Λ is given by (4.6). With respect to these spaces, continuity of all of these forms and the functional L are straightforward. The coercivity of $a(\cdot,\cdot)$ has been established, so it remains to verify that $b(\cdot,\cdot)$ and $d(\cdot,\cdot)$ are respectively V- and Λ-elliptic. The V-ellipticity of b is in fact trivial, and follows from the corresponding result for the elastic problem; while for d the desired result follows from a minor modification of earlier arguments.

The bounding scalar r_1 in (4.20) is easily estimated for the elastoplasticity problem. Assuming for convenience that the material is homogeneous so that the various material parameters are constants, it is readily shown that

$$r_1 \sim \frac{\lambda + 2\mu}{2\mu(1 + \min(k_1, k_2))}\,. \tag{4.35}$$

It follows that a sufficiently high degree of hardening and therefore a sufficiently large value of k_1 or k_2 would suffice to guarantee convergence of the algorithm.

The return map. The version of the algorithm most commonly found in applications, and which has been developed in various special forms, takes the form of a consistent tangent predictor step [26] followed by a corrector step that has a simple geometric interpretation, in the spaces of stresses, known as the return map [16, 25, 27]. The connection between this form of the corrector and that discussed in this work may be made using the local form (1.25) of the inequality for plastic flow. Assuming for convenience that $g = 0$ and $k_1 = 0$, the time-discrete version of this inequality, using a backward Euler approximation, states that at time t_n

$$\Delta \boldsymbol{p} \in N_{\mathcal{E}}(\boldsymbol{\sigma}_n)\,. \tag{4.36}$$

The stress at time t_n may be written as

$$\boldsymbol{\sigma}_n = \boldsymbol{\sigma}_n^{\mathrm{tr}} - \mathsf{C}\Delta \boldsymbol{p}_n\,, \tag{4.37}$$

where $\boldsymbol{\sigma}_n^{\mathrm{tr}} = \mathsf{C}[\boldsymbol{\varepsilon}(\boldsymbol{u}_n) - \boldsymbol{p}_{n-1}]$ is the trial elastic stress, that is, the stress at time t_n assuming that no plastic flow takes place in the time step $[t_{n-1}, t_n]$. Using (4.37) the inclusion (4.36) becomes, at time t_n,

$$\mathsf{C}^{-1}(\boldsymbol{\sigma}_n^{\mathrm{tr}} - \boldsymbol{\sigma}_n) \in N_{\mathcal{E}}(\boldsymbol{\sigma}_n)\,. \tag{4.38}$$

In other words, the actual stress $\boldsymbol{\sigma}_n$ may be obtained as the orthogonal projection, in the inner product generated by C^{-1}, of the trial elastic stress onto the convex elastic domain $\mathcal{E}$. This approach, proposed in an abstract framework in [19, 20], is referred to there as a catching-up algorithm or sweeping process.

The approach taken in this work has been to formulate and analyse predictor-corrector schemes with the corrector step based on the catching-up strategy, but to do so in the *equivalent* framework of the abstract variational inequality (4.1), which is well suited to analysis and to determining conditions for the entire algorithm to be convergent.

5. Concluding remarks

The focus of this work has been on the numerical analysis of a variational inequality motivated by a mathematical model of elastoplasticity. Results on the convergence of fully

discrete – that is, in space and time – approximations of the variational inequality have been summarized. A predictor-corrector algorithm, of the kind in common use in applications to elastoplasticity, has been presented and shown to be convergent under mild conditions on the data.

Rate-independent systems, of which the problem considered in this work is an example, have been given a comprehensive treatment in a framework developed by Mielke and coauthors (see for example [18] and the works cited in this survey paper). The essence of the framework is a weak formulation based on an energy balance equation and a stability inequality. Results on well-posedness of a broad range of rate-independent problems have been established by exploiting this energetic approach. Applications include perfect plasticity, which for the problem (1.21)–(1.30) corresponds to the case $k_1 = k_2 = 0$. In this case the displacement belongs to the space $BD(\Omega)$ of functions of bounded deformation, that is, integrable functions, whose symmetric gradients are bounded measures. The energetic method has been the basis for analysis of the more complex problem of large-deformation plasticity [18].

Both theoretical and numerical aspects of elastoplasticity and related mathematical problems continue to receive abundant attention. A major focus in recent years has been on strain-gradient theories, which are extensions of the model presented here, and which are appropriate models at the mesoscale, at which size effects are important. The extensions of the results reported here to the strain-gradient case are treated in [10], while optimal $O(h)$ convergence rates are proved in [23] for finite element approximations of a model of strain-gradient plasticity.

Acknowledgemenets. The author thanks the Department of Science and Technology and the National Research Foundation for their support through the South African Research Chair in Computational Mechanics.

References

[1] Carstensen, C., *Numerical Analysis of the Primal Problem of Elastoplasticity with Hardening*, Numer. Math. **82** (1999), 577–597.

[2] Ciarlet, P. G., *The Finite Element Method for Elliptic Problems*, SIAM, Philadelphia, 2002.

[3] Djoko, J. K., Ebobisse, F., McBride, A. T., and Reddy, B. D., *A Discontinuous Galerkin Formulation for Classical and Gradient Plasticity. Part 1: Formulation and Analysis*, Comp. Meths. Appl. Mech. Engng **196** (2007), 3881–3897.

[4] _____, *A Discontinuous Galerkin Formulation for Classical and Gradient Plasticity. Part 2: Algorithms and Numerical Analysis*, Comp. Meths. Appl. Mech. Engng **197** (2007), 1–21.

[5] Duvaut, G. and Lions, J.-L., *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.

[6] Fuchs, M. and Seregin, G., *Variational Methods for Problems from Plasticity Theory*

and for Generalized Newtonian Fluids, Lecture Notes in Mathematics 1749, Springer, Berlin, 2000.

[7] Glowinski, R. *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[8] Glowinski, R., Lions, J.-L., and Trémolières, R., *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[9] Han, W., Reddy, B. D., *Convergence of approximations to the primal problem in plasticity under conditions of minimal regularity*, Numer. Math. **87** (2000), 283–315.

[10] _____, *Plasticity: Mathematical Theory and Numerical Analysis*, Springer, New York. Second edition, 2013.

[11] Han, W., Reddy, B. D., and Schroeder, G. C., *Qualitative and Numerical Analysis of Quasistatic Problems in Elastoplasticity*, SIAM J. Numer. Anal. **34** (1997), 143–177.

[12] Kikuchi, N., Oden, J. T., *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM, Philadelphia, 1988.

[13] Kinderlehrer, D., Stampacchia, G., *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[14] D. Knees, *On global spatial regularity in elasto-plasticity with linear hardening*, Calc. Var. and PDEs **36** (2009), 611–625.

[15] _____, *On global spatial regularity and convergence rates for time dependent elasto-plasticity*, Math. Models Meths. Appl. Sci. **20** (2010), 1823–1858.

[16] Krieg, D. S. and Key, S. W., *Implementation of a Time Dependent Plasticity Theory into Structural Computer Programs*, in: Stricklin, J. A., Saczalski, K. J. (Eds.), Constitutive Equations in Viscoplasticity: Computational and Engineering Aspects, AMD-20, ASME, New York, 1976.

[17] Martin, J. B and Caddemi, S., *Sufficient Conditions for the Convergence of the Newton-Raphson Iterative Algorithm in Incremental Elastic-Plastic Analysis*, Euro. J. Mech. A/Solids **13** (1994), 351–365.

[18] Mielke, A., *Evolution in Rate-Independent Systems*, in Handbook of Differential Equations: Evolutionary Differential Equations, Dafermos, C., Feiereisl, E. (Eds.), Vol.2, Elsevier, 2005, 461–559.

[19] Moreau, J. J., *Evolution Problem Associated with a Moving Convex Set in a Hilbert Space*, J. Diff. Eqns **26** (1977), 347–374.

[20] _____, *Numerical Aspects of the Sweeping Process*, Comp. Meths. Appl. Mech. Engng **177** (1999), 329–349.

[21] Mosolov, P. P and Miasnikov, V. P., *Variational Nethods in the Theory of the Fluidity of a Viscous-Plastic Medium*, J. Appl. Math. Mech. **29** (1965), 545–577.

[22] Reddy, B. D., *Existence of Solutions to a Quasistatic Problem in Elastoplasticity*, in Bandle, C. et al. (Eds.), Progress in Partial Differential Equations: Calculus of Variations, Applications, Pitman Research Notes in Mathematics **267**, Longman, London, 1992, 233–259.

[23] Reddy, B. D., Wieners, C., and Wohlmuth, B., *Finite Element Analysis and Algorithms for Single-Crystal Strain-Gradient Plasticity*, Int. J. Numer. Meth. Engng **90** (2012), 784–804.

[24] Richtmyer, R. D. and Morton, K. W., *Difference Methods for Initial-Value Problems*, 2nd Edition, Interscience Pub.,. New-York, 1967.

[25] Simo, J. C., *Topics on the Numerical Analysis and Simulation of Plasticity*, in Ciarlet, P. G., Lions, J.-L. (Eds.), Handbook of Numerical Analysis, Vol. VI, North-Holland, Amsterdam, 1998, 183–499.

[26] Simo, J. C. and Taylor, R. L., *Consistent Tangent Operators for Rate-Independent Elasto-Plasticity*, Comp. Meth. Appl. Mech. Engng **48** (1985), 101–118.

[27] Wilkins, M.L., *Calculation of Elastic-Plastic Flow*, in Alder, B. (Ed.), Methods in Computational Physics, Vol. 3, Academic Press, New York, 1964, 211–263.

Department of Mathematics and Applied Mathematics, University of Cape Town, 7701 Rondebosch, South Africa

E-mail: daya.reddy@uct.ac.za

Uncertainty quantification in Bayesian inversion

Andrew M. Stuart

Abstract. Probabilistic thinking is of growing importance in many areas of mathematics. This paper highlights the beautiful mathematical framework, coupled with practical algorithms, which results from thinking probabilistically about inverse problems arising in partial differential equations.

Many inverse problems in the physical sciences require the determination of an unknown field from a finite set of indirect measurements. Examples include oceanography, oil recovery, water resource management and weather forecasting. In the Bayesian approach to these problems, the unknown and the data are modelled as a jointly varying random variable, typically linked through solution of a partial differential equation, and the solution of the inverse problem is the distribution of the unknown given the data.

This approach provides a natural way to provide estimates of the unknown field, together with a quantification of the uncertainty associated with the estimate. It is hence a useful practical modelling tool. However it also provides a very elegant mathematical framework for inverse problems: whilst the classical approach to inverse problems leads to ill-posedness, the Bayesian approach leads to a natural well-posedness and stability theory. Furthermore this framework provides a way of deriving and developing algorithms which are well-suited to the formidable computational challenges which arise from the conjunction of approximations arising from the numerical analysis of partial differential equations, together with approximations of central limit theorem type arising from sampling of measures.

Mathematics Subject Classification (2010). Primary 35R30; Secondary 62C10.

Keywords. Inverse problems, Bayesian inversion, uncertainty quantification, Monte Carlo methods, stochastic partial differential equations.

1. Introduction

Let X, R be Banach spaces and $G : X \to R$. For example G might represent the *forward* map which takes the input data $u \in X$ for a partial differential equation (PDE) into the solution $r \in R$. **Uncertainty quantification** is concerned with determining the propagation of randomness in the input u into randomness in some *quantity of interest* $q \in Q$, with Q again a Banach space, found by applying operator $\mathcal{Q} : R \to Q$ to $G(u)$; thus $q = (\mathcal{Q} \circ G)(u)$. The situation is illustrated in Figure 1.1.

Inverse problems are concerned with the related problem of determining the input u when given noisy *observed data* y found from $G(u)$. Let Y be the Banach space where the observations lie, let $\mathcal{O} : R \to Y$ denote the *observation operator*, define $\mathcal{G} = \mathcal{O} \circ G$, and consider the equation

$$y = \mathcal{G}(u) + \eta \tag{1.1}$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

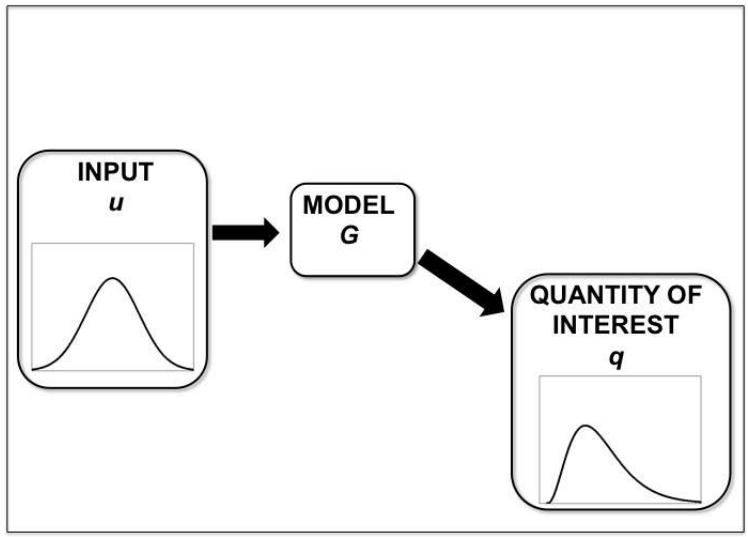


Figure 1.1. Uncertainty Quantification

viewed as an equation for $u \in X$ given $y \in Y$. The element $\eta \in Y$ represents *noise*, and typically something about the size of η is assumed known, often only in a statistical sense, but the actual instance of η entering the data y is not known. The aim is to reconstruct u from y. The **Bayesian inverse problem** is to find the the conditional probability distribution on $u|y$ from the joint distribution of the random variable (u, y); the latter is determined by specifying the distributions on u and η and, for example, assuming that u and η are independent. This situation is illustrated in Figure 1.2.

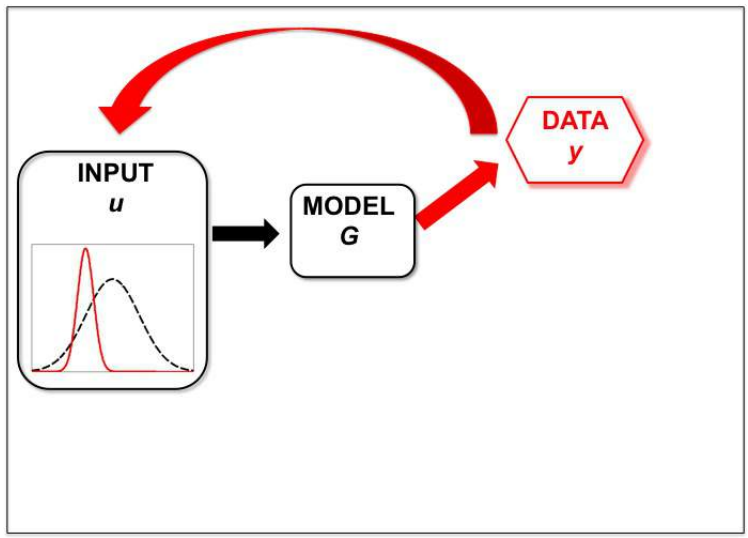


Figure 1.2. Bayesian Inverse Problem

To formulate the inverse problem probabilistically it is natural to work with separable Banach spaces as this allows for development of an integration theory (Bochner) as well as avoiding a variety of pathologies that might otherwise arise; we assume separability from now on. The probability measure on u is termed the *prior*, and will be denoted by μ_0, and that on $u|y$ the *posterior*, and will be denoted by μ^y. Once the Bayesian inverse problems has been solved, the uncertainty in q can be quantified with respect to input distributed according to the posterior on $u|y$, resulting in improved quantification of uncertainty in comparison with simply using input distributed according to the prior on u. The situation is illustrated in Figure 1.3. The black dotted lines demonstrate uncertainty quantification prior to incorporating the data, the red curves demonstrate uncertainty quantification after the data has been incoprorated by means of Bayesian inversion.

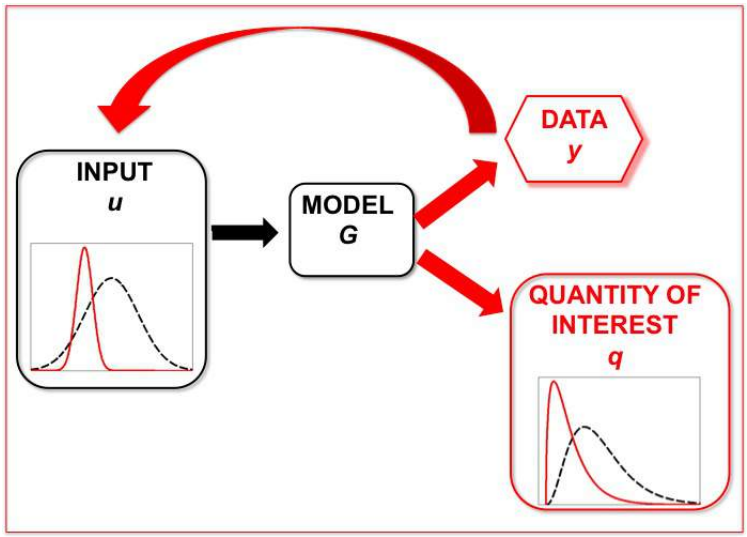


Figure 1.3. Uncertainty Quantification in Bayesian Inversion.

Carrying out the program illusrated in Figure 1.3 can have enormous benefits within a wide-range of important problems arising in science and technology. This is illustrated in Figure 1.4. The top two panels show representative draws from the prior (left) and posterior (right) probability distribution on the geological properties of a subsurface oil field, whilst the bottom two panels show predictions of future oil production, with uncertainty represented via the spread of the ensemble of outcomes shown, again under the prior on the left and under the posterior on the right. The unknown u here is the log permeability of the subsurface, the data y comprises measurements at oil wells and the quantity of interest q is future oil production. The map G is the solution of a system of partial differential equations (PDEs) describing the two-phase flow of oil-water in a porous medium, in which u enters as an unknown coefficient. The figure demonstrates that the use of data significantly reduces the uncertainty in the predictions.

The reader is hopefully persuaded, then, of the power of combining a mathematical model with data. Furthermore it should also be apparent that the set-up described applies to an enormous range of applications; it is also robust to changes, such as allowing for correlation between the noise η and the element $u \in X$. However, producing Figure 1.4, and similar

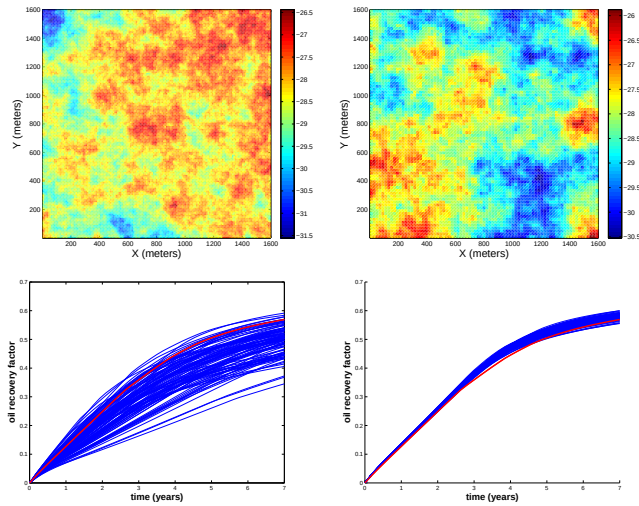


Figure 1.4. Upper panels: typical draws from the prior (left) and posterior (right). Lower panels: uncertainty in oil production under the prior (left) and posterior (right).

in other application areas, is a demanding computational task: it requires the full power of numerical analysis, to approximate the forward map G, and the full power of computational statistics, to probe the posterior distribution. The central thrust of the mathematical research which underlies this talk is concerned with how to undertake such tasks efficiently. The key idea underlying all of the work is to conceive of Bayesian inversion in the separable Banach space X, to conceive of algorithms for probing the measure μ^y on X and, only once this has been done, to then apply discretization of the unknown field u, to a finite dimensional space $\mathbb{R}^N$, and discretization of the forward PDE solver. This differs from a great deal of applied work which discretizes the space X at the very start to obtain a measure $\mu^{y,N}$ on $\mathbb{R}^N$, and then employs standard statistical techniques on $\mathbb{R}^N$. The idea is illustrated in Figure 1.5. Of course algorithms derived by the black route and the red route *can* lead to algorithms which coincide; however many of the algorithms derived via the the black route do not behave well under refinement of the approximation, $N \to \infty$, whilst those derived via the red route do since they are designed to work on X where $N = \infty$. Conceptual problem formulation and algorithm development via the red route is thus advocated.

This may all seem rather discursive, but a great deal of mathematical meat has gone into making precise theories which back-up the philosophy. The short space provided here is not enough to do justice to the mathematics and the reader is directed to [72] for details. Here we confine ourselves to a brief description of the historical context for the subject, given in section 2, and a summary of some of the novel mathematical and algorithmic ideas which have emerged to support the philisophy encapsulated in Figure 1.5, in sections 4 and 5. Section 3 contains some examples of inverse problems which motivated the theoretical work highlighted in sections 4 and 5, and may also serve to help the reader who prefers concrete settings. Section 6 contains some concluding remarks.

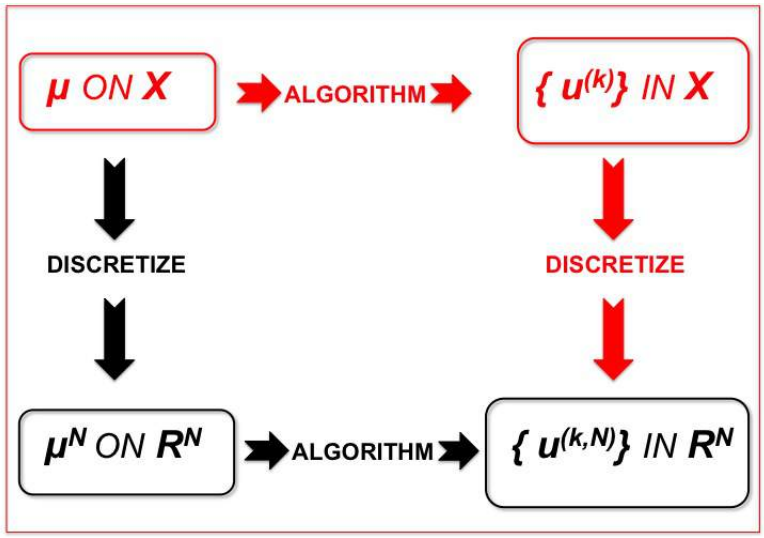


Figure 1.5. The red route is conceptually benefical in comparison with the black route.

2. Historical context

A cornerstone in the mathematical development of uncertainty quantification is the book [28] which unified and galvanized a growing engineering community interested in problems with random (uncertain) parameters. The next two and a half decades saw remarkable developments in this field, on both the applied and theoretical sides; in particular a systematic numerical analysis evolved which may be traced through the series of papers [5–7, 15–17, 58, 59, 61, 68, 76] and the referenes therein. Inverse problems have a long history and arise in an enormous range of applications and mathematical formulations. The 1976 article of Keller [37] is widely cited as foundational in the classical approach to inverse problems, and the modern classical theory, especially in relation to PDE and integral equations, is overviewed in a variety of texts: see [25, 38], for example.

The classical theory of inverse problems does not quantify uncertainty: typically it employs knowledge of the size of η but not its statistical distribution. However as long ago as 1970 the possibility of formulating PDE inverse problems in terms of Bayes' formula on the space X was recognized by Franklin [27] who studied classical linear inverse problems, such as inverting the heat kernel, from this perspective. That paper focussed on the rational basis for deriving a regularization using the Bayesian approach, rather than on quantifying uncertainty, but the posterior (Gaussian in this case) distribution did indeed provide a quantification of uncertainty. However it is arguable that the work of Franklin was so far ahead of its time that it made little impact when it appeared, primarily because the computational power needed to approach practical problems from this perspective was not available. The book of Kaipio and Somersalo [39] in 2005, however, had immediate impact, laying out a Bayesian methodology for inverse problems, and demonstrating its applicability to a range of important applications; computer power was ripe for the exploitation of fully Bayesian analyses when the book was published. However the perspective in [39] corresponded essentially to the black route outlined in Figure 1.5 ($N < \infty$) and did not take an infinite

dimensional perspective in X.

In the interim between 1970 and 2005 there had been significant development of the theory of Bayesian inversion in X for linear problems, building on the work of Franklin [49, 53], and working directly in the infinite dimensional space X. Lasanen then developed this into a fully nonlinear theory [44, 45, 47, 48], also working on X. This theoretical work was not concerned directly with the development of practical algorithms and the need to interface computational Bayesian practice with numerical analysis; in particular the need to deal with limits $N \to \infty$ in order to represent elements of X was not addressed. However others within the Bayesian school of inverse problems were interested in this question; see, for example, the paper [50]. Furthermore, in contrast to classical inversion, which is (often by definition [25]) ill-posed, Bayesian inversion comes with a desirable well-posedness theory on X which, itself, underpins approximation theories [71]; we will survey some of the developments which come from this perspective in what follows. Cousins of this well-posedness theory on X may be found in the papers [54, 57] both of which consider issues relating to perturbation of the posterior, in the finite dimensional setting $N < \infty$.

The primary applications which drive the theoretical and algorithmic developments highlighted in this article are in subsurface geophysics and in the atmosphere-ocean sciences. In the subsurface two major forces for the adoption of the Bayesian approach to inversion have been the work of Tarantola and co-workers and of Oliver and co-workers; see the books [60, 75] for further references. In the ocean-atmosphere sciences the Bayesian perspective has been less popular, but the book of Bennett [9] makes a strong case for it, primarily in the oceanographic context, whilst the work of Lorenc [52] has been a powerful force for Bayesian thinking in numerical weather prediction.

3. Examples

We provide in this section three examples to aid the reader who prefers concrete applications, and to highlight the type of problems which have motivated the theoretical develepments overviewed in the following sections. All of the examples can be placed in the general framework of (1.1).

3.1. Linear inverse problem. Consider the bounded linear map $K : X \to Y$, with X, Y separable Banach spaces, and the problem of finding $u \in X$ from noisy observations y of the image of u under K, given by

$$y = Ku + \eta.$$

For example if u is the initial condition of the heat equation on bounded open set $D \subset \mathbb{R}^d$, $X = L^2(D)$ and K denotes the solution operator for the heat equation over time interval T, then this is a widely used example of a classically ill-posed inverse problem. Ill-posedness arises beause of the smoothing property of the heat kernel and the fact that the noise η may take y out of the range space of K. Further ill-posedness can arise, for example, if K is found from the composition of the solution operator for the heat equation over time interval T with an operator comprising a finite set of point evaluations; the need to find a function u from a finite set of observations then leads to the problem being under-determined, further compounding ill-posedness. Linear inverse problems were the subject of the foundational paper [27], and developed further in [49, 53]. Natural applications include image processing.

3.2. Data assimilation in fluid mechanics. A natural nonlinear generalization of the inverse problem for the heat equation, and one which is prototypical of the inverse problems arising in oceanography and weather forecasting, is the following. Consider the Navier-Stokes equation written as an ordinary differential equation in the Hilbert space $X \coloneqq L^2_{\mathrm{div}}(\mathbb{T}^2)$ of square-integrable divergence-free functions on the two-dimensional torus:

$$\frac{dv}{dt} + \nu Av + B(v,v) = f, \quad v(0) = u \in X.$$

This describes the velocity field $v(x,t)$ for a model of incompressible Newtonian flow [73] on a two-dimensional periodic domain. An inverse problem protoypical of weather forecasting in particular is to find $u \in X$ given noisy *Eulerian observations*

$$y_{j,k} = v(x_j, t_k) + \eta_{j,k}.$$

Like the heat equation the forward solution operator is smoothing, and the fact that the observations are finite in number further compounds ill-posedness. In addition the nonlinearity adds further complications, such as sensitive dependence on initial conditions arising from the chaotic character of the equations for $\nu \ll 1$. There are many interesting variants on this problem; one is to consider *Lagrangian observations* derived from tracers moving according to the velocity field v itself, and the problem is prototypical of inverse problems which arise in oceanography. Determining the initial condition of models from fluid mechanics on the basis of observations at later times is termed *data assimilation*. Both Eulerian and Lagrangian data assimilation are formulated as Bayesain inverse problems in [13].

3.3. Groundwater flow. The following is prototypical of inverse problems arising in hydrology and in oil reservoir modelling. Consider the Darcy Flow, with log permeability $u \in X = L^\infty(D)$, described by the equation

$$\begin{aligned} -\nabla \cdot \big(\exp(u)\nabla p\big) &= 0, \quad x \in D, \\ u &= g, \quad x \in \partial D. \end{aligned}$$

Here the aim is to find $u \in X$ given noisy observations

$$y_j = p(x_j) + \eta_j.$$

The pressure p is a surrogate for the height of the water table and measurements of this height are made by hydrologists seeking to understand the earths subsurface. The resulting classical inverse problem is studied in [66] and Bayesian formulations are given in [21, 22]. The space $L^\infty(D)$ is not separable, but this difficulty can be circumvented by working in separable Banach spaces found as the closure of the linear span of an infinite set of functions in $L^\infty(D)$, with respect to the $L^\infty(D)$-norm.

4. Mathematical foundations

In this section we briefly outline some of the issues involved in the rigorous formulation of Bayesian inversion on a separable Banach space X. We start by discussing various prior models on X, and then discuss how Bayes' formula may be used to incorporate data and update these prior distributions on u into posterior distributions on $u|y$.

4.1. Priors: Random functions. Perhaps the simplest way to construct random priors on a function space X is as follows. Let $\{\varphi_j\}_{j=1}^{\infty}$ denotes an infinite sequence in the Banach space X, normalized so that $\|\varphi_j\|_X = 1$. Define the deterministic random sequence $\gamma = \{\gamma_j\}_{j=1}^{\infty} \in \ell_w^p(\mathbb{R})$, where $\ell_w^p(\mathbb{R})$ denotes the sequence of $p^{th}-$power summable sequences, when weighted by the sequence $w = \{w_j\}_{j=1}^{\infty}$. Then let $\xi = \{\xi_j\}_{j=1}^{\infty}$ denote the i.i.d sequence of centred random variables in $\mathbb{R}$, normalized to that $\mathbb{E}\xi_1^2 = 1$. We define $u_j = \gamma_j\xi_j$ and pick a *mean* element $m \in X$ and then consider the random function

$$u = m + \sum_{j=1}^{\infty} u_j\varphi_j. \tag{4.1}$$

The probability measure on the random sequence implies, via its pushforward under the construction (4.1) a probability measure on the function u; we denote this measure by μ_0. Of course the fact that the φ_j are elements of X does not imply that μ_0 is a measure on X: assumptions must be made on the decay of the sequence γ. For example, using the fact that the random sequence $u = \{u_j\}_{j=1}^{\infty}$ comprises independent centred random variables we find that

$$\mathbb{E}^{\mu_0}\|u-m\|_X^2 = \sum_{j=1}^{\infty}\gamma_j^2.$$

This demonstrates that assuming $\gamma = \{\gamma_j\}_{j=1}^{\infty} \in \ell^2(\mathbb{R})$ is sufficient to ensure that the random function is almost surely an element of X. If the space X itself is not separable, this difficulty can be circumvented by working in a separable Banach space X' found as the closure of the linear span of the φ_j with respect to the norm in X.

Expansions of the form (4.1) go by the name Karhunen-Loeve in the Gaussian case [1] arising when ξ_1 is a Gaussian random variable. The so-called Besov case was introduced in [50] and concerns the case where ξ_1 is distributed according to Lebesgue density proportional to a power of $\exp(-|\cdot|^q)$, subsuming the Gaussian situation as the special case $q = 2$. Schwab has been a leading proponent of random functions constructed using compactly supported random variables ξ_1 – see [68, 70] and the references therein; although not so natural from an applications viewpoint, the simplicity that follows from this assumption allows the study of key issues in uncertainty quantification and Bayesian inversion without the need to deal with a variety of substantial technical issues which arise when ξ_1 is not compactly supported; in particular integrability of the tails becomes a key technical issue for non-compactly supported ξ_1, and there is a need for a Fernique theorem [26] or its analogue [22, 50]. For a general treatment of random functions constructed as in (4.1) see the book Kahane [36].

4.2. Priors: Hierarchical. There are many parameters required in the prior constructions of the previous subsection, and in many applications these may not be known. In such situations these parameters can be inferred from the data, along with u. Rather than giving a general discussion we consider the example of Gaussian priors when X is a Hilbert space. A draw u from a Gaussian is written as $u \sim N(m, C)$ where $N(m, C)$ denotes a Gaussian with mean m and covariance C. Here the covariance operator C is defined by

$$\begin{aligned} C &= \mathbb{E}^{\mu_0}(u-m)\otimes(u-m) \\ &= \sum_{j=1}^{\infty}\gamma_j^2\varphi_j\otimes\varphi_j. \end{aligned}$$

Note that then

$$C\varphi_j = \gamma_j^2 \varphi_j.$$

An example hierarchical prior may be constructed by introducing an unknown parameter δ, which scales the covariance, and positing that

$$u|\delta \sim N(0, \delta^{-1}C).$$
$$\delta \sim \mathrm{Ga}(\alpha, \beta).$$

Here Ga denotes the Gamma distribution, and of course other prior assumptions on δ are possible. The potential for the use of hierarchical priors in linear inverse problems has been highlighted in several recent papers, see [8, 10, 11] for example, all in the finite dimensional context; such models have been studied in the large dimension and infinite dimensional limit in [2].

4.3. Priors: Geometric. The probability measures constructed through random functions are inherently infinite dimensional, being built on an infinite sequence of random coefficients. In the previous subsection we showed how these could be extended to priors which included an extra unknown parameter δ specifying the scale of the prior; there are numerous generalizations of this basic concept. Here we describe one of them that is particularly useful in the study of subsurface inverse problems where the geometry imposed by faults, old fluival structures and so forth is a major determining fact in underground porous medium fluid flow.

Examples of problems to which our theory applies may be found in Figure 4.1. In the top left we show a layered structure in which a piecewise constant function is constructed; this maybe generalized to include faults, as in the bottom left. The top right shows a generalization of the layered structured to allow a different Gaussian random field realization in each layer, and the bottom right shows a generalization to allow for a channel-like structure, typical of fluvial deposition.

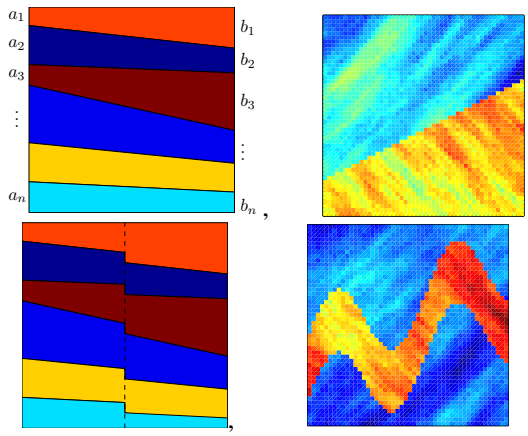


Figure 4.1. Uncertainty quantification under the prior and the posterior

.

The development of layered prior models was pioneered in [12]. The chanellized structure as prior was developed in [43] and [78]. All of this work was finite dimensional, but a

theoretical frameowork subsuming these particular cases, and set in infinite dimensions, is developed in [35].

4.4. Posterior. Recall that the Bayesuah solutiuon to the inverse problem of funding u from data y given by (1.1) is to determine the probability distribution on $u|y$, which lives on the space X, from the probability distribution of the joint random variable (u, y) which lives on $X \times Y$. In order to do this we specify to the situation where $Y = \mathbb{R}^J$, so that the number of observations is finite, and assume that $\eta \sim N(0, \Gamma)$, with Γ an invertible covariance matrix on $\mathbb{R}^J$. Many generalizations of this are possible, to both infinite dimensions or to non-Gaussian noise η, but the setting with fnite dimensional data allows us to expose the main ideas.

We define the *model-data mismatch functional*, or *least squares functional*, given by

$$\Phi(u; y) := \frac{1}{2}\big|\Gamma^{-\frac{1}{2}}\big(y - \mathcal{G}(u)\big)\big|^2$$

where $|\cdot|$ denotes the Euclidean norm. Classical Bayesian inversion is concerned with minimizing $\Phi(\cdot; y)$, typically with incoporation of regularization through addition of a penalty term (Tikhonov regularization) or through specification of seeking minimizers within a compact subset of X [25]. It is natural to ask how a Bayesian approach relates to such classical approaches.

Bayes' formula is typically stated as

$$\frac{\mathbb{P}(u|y)}{\mathbb{P}(u)} \propto \mathbb{P}(y|u)$$

and our wish is to formulate this precisely in the infinite dimensional context where u lives in a separable Banach space. Given a prior measure μ_0 on u and a posterior measure μ^y on $u|y$ a typical infinite dimensional version of Bayes' formula is a statement that μ^y is absolutely continuous with respect to μ_0 and that

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\Big(-\Phi(u; y)\Big). \tag{4.2}$$

Note that the righ-hand side is indeed proportional to $\mathbb{P}(y|u)$ whilst the left-hand side is an infinite dimensional analogue of $\frac{\mathbb{P}(u|y)}{\mathbb{P}(u)}$. The formula (4.2) implies that the posterior measure is large (resp. small), relative to the prior measure, on sets where $\Phi(\cdot; y)$ is small (resp. large). As such we see a clear link between classical inversion, which aims to choose elements of X which make $\Phi(\cdot; y)$ small, and the Bayesian approach.

There is a particular structure which occurs in the linear inverse problem of subsection 3.1, namely that if η is distributed according to a Gaussian, then the posterior on $u|y$ is Gaussian if the prior on u is Gaussian; the prior and posterior are termed *conjugate* in this situation, coming from the same class. See [3, 41] for a discussion of this Gaussian conjugacy for linear inverse problems in infinite dimensions.

4.5. Well-Posed posterior. For a wide range of the priors and examples given previously there is a well-posedness theory which accompanies the Bayesian perspective. This theory is developed, for example, in the papers [13, 21, 22, 35, 71]. This theory shows that the posterior μ^y is Hölder in the Hellinger metric with respect to changes in the data y. The

Hölder exponent depends on the prior, and is one (the Lipschitz case) for many applications. However it is important to strike a note of caution concerning the robustness of the Bayesian approach: see [62].

4.6. Recovery of truth. Consider data y given from truth $\mathrm{u}^\dagger$ by

$$y = \mathcal{G}(\mathrm{u}^\dagger) + \epsilon\,\eta_0, \quad \eta_0 \sim N(0, \Gamma_0).$$

Thus we have assumed that the data is generated from the model used to construct the posterior. It is then natural to ask how close is the posterior measure μ^y to the truth $\mathrm{u}^\dagger$? For many of the preceding problems we have (refinements of) results of the type:

$$\text{For any } \delta > 0,\ \mathbb{P}^{\mu^y}\left(|u - \mathrm{u}^\dagger| > \delta\right) \to 0 \text{ as } \epsilon \to 0.$$

Examples of theories of this type may be found for linear problems of subsection 3.1 in [3, 4, 41, 42, 46, 65], for the Eulerian Navier-Stokes inverse problems of subsection 3.2 in [67], and for the groundwater flow problem of subsection 3.3 in [77].

5. Algorithms

The preceding chapter describes a range of theoretical developments which allow for precise characterizations of, and study of the properties of, the posterior distribution μ^y. These are interesting in their own right, but they also underpin algorithmic approaches which aim to be efficient with respect to increase of N in the approximation of μ^y by a measure $\mu^{y,N}$ on $\mathbb{R}^N$. Here we outline research in this direction.

5.1. Forward error = Inverse error. Imagine that we have approximated the space X by $\mathbb{R}^N$; for example we might truncate the expansion (4.1) at N terms and consider the inverse problem for the N unknown coefficients in the representation of u. We then approximate the forward map $\mathcal{G}$ by a numerical method to obtain $\mathcal{G}^N$ satisfying, for u in X,

$$|\mathcal{G}(u) - \mathcal{G}^N(u)| \le \psi(N) \to 0$$

as $N \to \infty$. Such results are in the domain of classical numerical analysis. It is interesting to understand their implications for the Bayesian inverse problem.

The approximation of the forward map leads to an approximate posterior measure $\mu^{y,N}$ and it is natural to ask how expectations under μ^y, the ideal expectations to be computed, and under $\mu^{y,N}$, expectations under which we may approximate by, for example statistical sampling techniques, compare. Under quite general conditions it is possible to prove [18] that, for an appropriate class of test functions $f : X \to \mathsf{S}$, with S a Banach space,

$$\|\mathbb{E}^{\mu^y} f(u) - \mathbb{E}^{\mu^{y,N}} f(u)\|_{\mathsf{S}} \le C\psi(N).$$

The method used is to employ the stability in the Hellinger metric implied by the well-posedness theory to show that μ^y and $\mu^{y,N}$ are $\psi(\mathcal{N})$ close in the Hellinger metric and then use properties of that metric to bound perturbations in expectations.

5.2. Faster MCMC. The preceding subsection demonstrates how to control errors arising from the numerical analysis component of any approximation of a Bayesian inverse problem.

Here we turn to statistical sampling error, and in particular to Markov Chain-Monte Carlo (MCMC) methods. These methods were developed in the statistical physics community in [56] and then generalized to a flexible tool for statistical sampling in [33]. The paper [74] demonstrated an abstract framework for such methods on infinite dimensional spaces.

The full power of using MCMC methodology for inverse problems was highlighted in [39] and used for interesting applications in the subsurface in, for example, [24]. However for a wide range of priors/model problems it is possible to show that standard MCMC algorithms, derived by the black route in Figure 1.5, mix in $\mathcal{O}(N^a)$ steps, for some $a > 0$ implying undesirable slowing down as N increases. By following the red route in Figure 1.5, however, it is possible to create new MCMC algorithms which mix in $\mathcal{O}(1)$ steps.

The slowing down of standard MCMC methods in high dimensions is demonstrated by means of diffusion limits in [55] for Gaussian priors and in [2] for hierarchical Gaussian priors. Diffusion limits where then used to demonstrate the effectiveness of the new method, derived via the red route in Figure 1.5, in [63] and a review explaining the derivation of such new methods maybe found in [19]. The paper [31] uses spectral gaps to both quantify the benefits of the method studied in [63] ($\mathcal{O}(1)$ lower bounds on the spectral gap) compared with the drawbacks of traditional methods, such as that studied in [55] ($\mathcal{O}(N^{-\frac{1}{2}})$ upper bounds on the spectral gap.)

These new MCMC methods are starting to find their way into use within large-scale engineering inverse problems and to be extended and modified to make them more efficient in large data sets, or small noise data sets scenarios; see for examples [14, 20, 29].

5.3. Other directions. The previous subsection concentrated on a particular class of methods for exploring the posterior distribution, namely MCMC methods. These are by no means the only class of methods available for probing the posterior and here we give a brief overview of some other approaches that may be used.

The determinsitic approximation of posterior expectations, by means of sparse approximation of high dimensional integrals, is one approach with great potential. The mathematical theory behind this subject is overviewed in [68] in the context of standard uncertainty quantification, and the approach is extended to Bayesian inverse problems and uncertainty quantification in [70], with recent computational and theoretical progress contained in [69].

It is also possible to combine sparse approximation techniques with MCMC and the computational complexity of this approach is analyzed in [32], and references to the engineering literature, where this approach was pioneered, are given. The idea of multilevel Monte Carlo [30] has recently been generalized to MCMC methods; see the paper [32] which analyzes the computational complexity of such methods, and the paper [40] in which a variant on such methods was introduced and implemented for the groundwater flow problem.

Another computational approach, widely used in machine learning when complex probability measures need to be probed, is to look for the best approximation of μ^y within some simple class of measures. If the class comprises Dirac measures then such an approach is known as *maximum a posterior estimation* and corresponds in finite dimensions, when the posterior has a Lebesgue density, to finding the location of the peak of that density [39]. This idea is extended to the infinite dimensional setting in [23]. In the context of uncertainty quantification the MAP estimator itself is not of direct use as it contains no information about fluctuations. However linearization about the MAP can be used to compute a Gaussian approximation at that point. A more sophisticated approach is to directly seek the best Gaussian approximation $\nu = N(m, C)$ wrt relative entropy. Analysis of this in the infinite

dimensional setting, viewed as a problem in the calculus of variations, is undertaken in [64].

6. Conclusions

Combining uncertainty quantification with Bayesian inversion provides formidable computational challenges relating to the need to control, and optimally balance, errors arising from the numerical analysis, and approximation of the forward operator, with errors arising from computational statistical probing of the posterior distribution. The approach to this problem outlined here has been to adopt a way of deriving and analyzing algorithms based on thinking about them in infinite dimensional spaces, and only then discretizing to obtain implementable algorithms in $\mathbb{R}^N$ with $N < \infty$. This requires formulation and analysis of the Bayesian inverse problem in infinite dimensions. We have overviewed the mathematical theory that goes into this formulation and analysis, in section 3, and overviewed the algorithmic developments which follow from it, in section 4.

In some applications it is starting to be feasible to compute accurate approximations of the Bayesian posterior distribution, and it is to be expected that there will be great strides in this area over the next decade, both in terms of range of applications and algorithmic innovation, with the latter based on the infinite dimensional perspective given here, but making more careful exploitation of data and structure of the likelihood. Even where the fully Bayesian approach is out of the question for the forseeable future, for example in weather forecasting, the Bayesian approach described here can be important as it may be used as a gold standard against which to benchmark algorithms which are useable in practice. This approach is employed in [34, 51] in the context of model problems of the type shown in sections 3.2 and 3.3, and variants on them.

Finally the reader is reminded that this article is in essay form and contains no mathematical details. For an overview of the subject in which mathematical details are given the reader is refered to [72].

Acknowledgements. The author is grateful to EPSRC, ERC and ONR for financial support which led to the work described in this lecture. He is grateful to Marco Iglesias for help in preparing the figures, and to Yuan-Xiang Zhang for careful proof-reading of the article.

References

[1] R. Adler, *The Geometry Of Random Fields*, SIAM, 1981.

[2] S. Agapiou, J. Bardsley, O. Papaspiliopoulis, and A.M. Stuart, *Analysis of the Gibbs sampler for hierarchical inverse problems*, arXiv:1311.1138.pdf

[3] S. Agapiou, S. Larsson, and A. M. Stuart, *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, Stochastic Processes and their Applications **123** (2013), 3828–3860.

[4] S. Agapiou, A. M. Stuart, and Y. X. Zhang, *Bayesian posterior consistency for linear severely ill-posed inverse problems*, To appear Journal of Inverse and Ill-posed Problems, arXiv:1210.1563

[5] I. Babuska, R. Tempone and G. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations.,* SIAM J. Num. Anal. **42** (2004), 800–825.

[6] _____, *Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation,* Applied Mechanics and Engineering, **194** (2005), 1251–1294.

[7] I. Babuska, F. Nobile and R. Tempone, *A Stochastic Collocation method for elliptic Partial Differential Equations with Random Input Data,* SIAM J. Num. Anal. **45** (2007), 1005–1034.

[8] J Bardsley, *MCMC-Based image reconstruction with uncertainty quantification*, SISC **34** (2012), 1316–1332.

[9] AF Bennett, *Inverse Modeling of the Ocean and Atmosphere*, Cambridge University Press, 2002.

[10] D. Calvetti, H.Hakula, S. Pursiainen, and E. Somersalo, *Conditionally Gaussian Hypermodels for Cerebral Source Localization*, SIAM J. Imaging Sciences **2** 2009, 879–909.

[11] D. Calvetti and E. Somersalo. *Hypermodels in the Bayesian imaging framework*, Inverse Problems **24** (2008), 034013.

[12] J. Carter and D. White, *History matching on the Imperial College fault model using parallel tempering*, Computational Geosciences **17** (2013), 43–65.

[13] S. Cotter, M. Dashti, J. Robinson, and A. Stuart, *Bayesian inverse problems for functions and applications to fluid mechanics*, Inverse Problems **25** (2009), doi:10.1088/0266–5611/25/11/115008.

[14] A. Cliffe, O. Ernst, and B. Sprungk, In preparation, 2014.

[15] A. Cohen, R. DeVore, and S. Schwab, *Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs*, Foundations of Computational Mathematics **10** (2010), 615–646.

[16] _____, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs*, Analysis and Applications **9** (2011), 11–47.

[17] A. Chkifa, A. Cohen, R. DeVore, and S. Schwab, *Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs*, ESAIM: Mathematical Modelling and Numerical Analysis **47** (2013), 253–280.

[18] S. Cotter, M. Dashti, and A. Stuart, *Approximation of Bayesian inverse problems*, SIAM Journal of Numerical Analysis **48** (2010), 322–345.

[19] S. Cotter, G. Roberts, A. Stuart, and D. White. *MCMC methods for functions: modifying old algorithms to make them faster*, Statistical Science, to appear, arXiv:1202.0709.

[20] T. Cui, K.J.H. Law and Y. Marzouk, In preparation, 2014.

[21] M. Dashti and A. Stuart, *Uncertainty quantification and weak approximation of an elliptic inverse problem*, SIAM J. Num. Anal. **49** (2012), 2524–2542.

[22] M. Dashti, S. Harris, and A. Stuart, *Besov priors for Bayesian inverse problems*, Inverse Problems and Imaging **6** (2012), 183–200.

[23] M. Dashti, KJH. Law, AM Stuart, and J.Voss, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems **29** (2013), 095017.

[24] P. Dostert, Y. Efendiev, T.Y. Hou, and W. Luo, *Coarse-gradient Langevin algorithms for dynamic data integration and uncertainty quantification*, Journal of Computational Physics **217** (2006), 123–142.

[25] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, 1996.

[26] X. Fernique. *Intégrabilité des vecteurs Gaussiens*, C. R. Acad. Sci. Paris Sér. A-B **270**, (1970), A1698–A1699.

[27] J. Franklin, *Well-posed stochastic extensions of ill-posed linear problems*, J. Math. Anal. Appl. **31** (1970), 682–716.

[28] RG Ghanem and PD Spanos, *Stochastic Finite Elements: a Spectral Approach*, newblock Springer, 1991.

[29] O. Ghattas and T. Bui-Thanh, *An Analysis of Infinite Dimensional Bayesian Inverse Shape Acoustic Scattering and its Numerical Approximation*, SIAM Journal on Uncertainty Quantification, Submitted, 2012.

[30] M.Giles *Multilevel Monte Carlo path simulation*, Operations Research **56** (2008), 607–617.

[31] M. Hairer, A.M. Stuart, and S. Vollmer, *Spectral Gaps for a Metropolis-Hastings Algorithm in Infinite Dimensions*, To appear, Ann. Appl. Prob. 2014. arXiv:1112.1392

[32] V.H. Hoang, C. Schwab, and A.M. Stuart. *Complexity analysis of accelerated MCMC methods for Bayesian inversion*, Inverse Problems **29** (2013), 085010.

[33] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika **57**, no. 1, (1970), 97–109.

[34] M. Iglesias, K.J.H. Law, and A.M. Stuart, *Evaluation of Gaussian approximations for data assimilation in reservoir models*, Computational Geosciences **17** (2013), 851–885.

[35] M. Iglesias, K. Lin, and A.M. Stuart, *Well-Posed Bayesian Geometric Inverse Problems Arising in Subsurface Flow*, arXiv:1401.5571.

[36] J.-P. Kahane, *Some random series of functions*, vol. 5 of Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 1985.

[37] J.B. Keller, *Inverse problems*, Am. Math. Mon. **83** (1976), 107–118.

[38] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer, 1996.

[39] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, vol. 160 of Applied Mathematical Sciences, Springer-Verlag, New York, 2005.

[40] C. Ketelsen, R. Scheichl, and Teckentrup, *A Hierarchical Multilevel Markov Chain Monte Carlo Algorithm with Applications to Uncertainty Quantification in Subsurface Flow*, arXiv:1303.7343

[41] B. Knapik, A. van Der Vaart, and J. van Zanten, *Bayesian inverse problems with Gaussian priors*, Ann. Statist. **39**, no. 5, (2011), 2626–2657.

[42] _____, *Bayesian recovery of the initial condition for the heat equation*, arXiv:1111.5876.

[43] J.L. Landa and R.N. Horne, *A procedure to integrate well test data, reservoir performance history and 4-D seismic information into a reservoir description*, SPE Annual Technical Conference 1997, 99–114.

[44] S. Lasanen, *Discretizations of generalized random variables with applications to inverse problems*, Ann. Acad. Sci. Fenn. Math. Diss., University of Oulu **130**.

[45] _____, *Measurements and infinite-dimensional statistical inverse theory*, PAMM **7**, (2007), 1080101–1080102.

[46] _____, *Posterior convergence for approximated unknowns in non-Gaussian statistical inverse problems*, Arxiv preprint arXiv:1112.0906.

[47] _____, *Non-Gaussian statistical inverse problems. Part I: Posterior distributions*, Inverse Problems and Imaging **6**, no. 2, (2012), 215–266.

[48] _____, *Non-Gaussian statistical inverse problems. Part II: Posterior distributions*, Inverse Problems and Imaging **6**, no. 2, (2012), 267–287.

[49] M. S. Lehtinen, L. Päivärinta, and E. Somersalo, *Linear inverse problems for generalised random variables*, Inverse Problems **5**, no. 4, (1989), 599–612, http://stacks.iop.org/0266-5611/5/599.

[50] M. Lassas, E. Saksman, and S. Siltanen, *Discretization-invariant Bayesian inversion and Besov space priors*, Inverse Problems and Imaging **3** (2009), 87–122.

[51] K.J.H. Law and A.M. Stuart, *Evaluating data assimilation algorithms*, Monthly Weather Review **140** (2012), 3757–3782.

[52] A.C. Lorenc and O. Hammon, *Objective quality control of observations using Bayesian methods. Theory, and a practical implementation*, Quarterly Journal of the Royal Meteorological Society, **114** (1988), 515–543.

[53] A. Mandelbaum, *Linear estimators and measurable linear transformations on a Hilbert space*, Z. Wahrsch. Verw. Gebiete **65**, no. 3, (1984), 385–397, http://dx.doi.org/10.1007/BF00533743.

[54] Y. Marzouk and D. Xiu, *A stochastic collocation approach to Bayesian inference in inverse problems*, Communications in Computational Physicss **6** (2009), 826–847.

[55] J. Mattingly, N. Pillai, and A. Stuart, *Diffusion limits of the random walk Metropolis algorithm in high dimensions*, Ann. Appl. Prob. **22** (2012), 881–930.

[56] N. Metropolis, R. Rosenbluth, M. Teller, and E. Teller, E*quations of state calculations by fast computing machines*, J. Chem. Phys. **21** (1953), 1087–1092.

[57] A. Mondal, Y. Efendiev, B. Mallicka, and A. Datta-Gupta, *Bayesian uncertainty quantification for flows in heterogeneous porous media using reversible jump Markov chain Monte Carlo methods*, Advances in Water Resources, **3** (2010), 241–256.

[58] F. Nobile, R. Tempone, and CG Webester, *A sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM Journal on Numerical Analysis, **46** (2008), 2309–2345.

[59] _____, *An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM Journal on Numerical Analysis **46** (2008), 2441–2442.

[60] D.S. Oliver, A.C. Reynolds, and N. Liu, *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, Cambridge University Press, 2008.

[61] H. Owhadi, C. Scovel, T.J. Sullivan, M. McKerns, and M. Ortiz, *Optimal Uncertainty Quantification*, SIAM Review **55** (2013), 271–345.

[62] H. Owhadi, C. Scovel, and T.J. Sullivan, *When Bayesian Inference Shatters*, arXiv: 1308.6306

[63] N. Pillai, A. Stuart, and A. Thiery, *Gradient flow from a random walk in Hilbert space*, To appear, Stochastic Partial Differential Equations, arXiv:1108.1494.

[64] F. Pinski, G. Simpson, A.M. Stuart, and H. Weber, *Kullback-Leibler Approximation for Probability Measures on Infinite Dimensional Spaces*, arXiv:1310.7845

[65] K. Ray, *Bayesian inverse problems with non-conjugate priors*, Electronic Journal of Statistics **7** (2013), 1–3169.

[66] G. Richter, *An inverse problem for the steady state diffusion equation*, SIAM Journal on Applied Mathematics **41**, no. 2, (1981), 210–221.

[67] D. Sanz-Alonso and A.M. Stuart, In preparation, 2014.

[68] C. Schwab and C.J. Gittelson, *Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs*, Acta Numer. **20** (2011).

[69] C. Schillings and C. Schwab, *Sparse, adaptive Smolyak quadratures for Bayesian inverse problems* , Inverse Problems **29** (2013), 065011.

[70] C. Schwab and A. Stuart, *Sparse deterministic approximation of Bayesian inverse problems*, Inverse Problems **28** (2012), 045003.

[71] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer. **19** (2010), 451–559.

[72] ———, *The Bayesian approach to inverse problems,* arXiv:1302.6989

[73] R. Temam, *Navier-Stokes equations*, AMS Chelsea Publishing, Providence, RI, 2001.

[74] L. Tierney, *A note on Metropolis-Hastings kernels for general state spaces*, Ann. Appl. Probab. **8**, no. 1, (1998), 1–9.

[75] A. Tarantola, *Inverse Problem Theory*, SIAM, 2005.

[76] R.Tempone, *Numerical Complexity Analysis of Weak Approximation of Stochastic Differential Equations*, PhD Thesis, KTH Stockholm, Sweden, 2002. http://www.nada.kth.se/utbildning/forsk.utb/avhandlingar/dokt/Tempone.pdf

[77] S. Vollmer, *Posterior consistency for Bayesian inverse problems through stability and regression results*, Inverse Problems **29** (2013), 125011.

[78] J. Xie, Y. Efendiev, and Datta-Gupta, *Uncertainty quantification in history matching of channelized reservoirs using Markov chain level set approaches*, SPE Reservoir Simulation Symposium, 2011.

Mathematics Institute, Warwick University, Coventry CV4 7AL, UK
E-mail: a.m.stuart@warwick.ac.uk

Stochastic modeling and methods in optimal portfolio construction

Dedicated to my father George Zariphopoulos (1930-2014)

Thaleia Zariphopoulou

Abstract. Optimal portfolio construction is one of the most fundamental problems in financial mathematics. The foundations of investment theory are discussed together with modeling issues and various methods for the analysis of the associated stochastic optimization problems. Among others, the classical expected utility and its robust extension are presented as well as the recently developed approach of forward investment performance. The mathematical tools come from stochastic optimization for controlled diffusions, duality and stochastic partial differential equations. Connections between the academic research and the investment practice are also discussed and, in particular, the challenges of reconciling normative and descriptive approaches.

Mathematics Subject Classification (2010). Primary 97M30; Secondary 91G80.

Keywords. expected utility, forward performance, stochastic PDE, robustness, duality, HJB equation, stochastic optimization, portfolio choice.

1. Introduction

Financial mathematics is a burgeoning area of research on the crossroads of stochastic processes, stochastic analysis, optimization, partial differential equations, finance, econometrics, statistics and financial economics. There are two main directions in the field related, respectively, to the so-called *sell* and *buy* sides of financial markets. The former deals with derivative valuation, hedging and risk management while the latter with investments and fund management.

Derivatives are financial contracts written on primary financial assets. Their development started in the late 1970s with the revolutionary idea of Black, Merton and Scholes of pricing via "perfect replication" of the derivatives' payoffs. Subsequently, the universal theory of arbitrage-free valuation, developed by Kreps, and Harrison and Pliska, was built on a surprising fit between stochastic calculus and quantitative needs. It revolutionized the derivatives industry, but its impact did not stop there. Because the theory provided a model-free approach to price and manage risks, the option pricing methodology has been applied in an array of applications, among others, corporate and non-corporate agreements, pension funds, government loan guarantees and insurance plans. In a different direction, applications of the theory resulted in a substantial growth of the fields of real options and decision analysis. Complex issues related, for example, to operational efficiency, financial flexibility,

Proceedings of the International Congress of Mathematicians, Seoul, 2014

contracting, and initiation and execution of R&D projects were revisited and analyzed using derivative valuation arguments. For the last three decades, the theoretical developments, technological advances, modeling innovations and creation of new derivatives products have been developing at a remarkable rate. The recent financial crisis cast a lot of blame upon derivatives and quantitative methods and, more generally, upon financial mathematics. Despite the heated debate on what went really wrong, the theory of derivatives remains one of the best examples of a perfect match among mathematical innovation, technological sophistication and direct real world applicability.

In the complementary side of finance practice, investments deal with capital allocation under market uncertainty. The objective is not to eliminate the inherent market risks - as it is the case with derivatives - but to exploit optimally the market opportunities while undertaking such risks. The overall goal is to assess the trade-off between risks and payoffs. For this, one needs to have, from the one hand, models that predict satisfactorily future asset prices and, from the other, mechanisms that measure in a practically meaningful way the performance of investment strategies. There are great challenges in both these directions. Estimating the drift of stock prices is a notoriously difficult problem. Moreover, building appropriate investment criteria that reflect the investors' attitude is extremely complex, for these criteria need to capture an array of human sentiments like risk aversion, loss aversion, ambiguity, prudence, impatience, etc.. There is extensive academic work, based on the foundational concept of expected utility, that examines such issues. However, there is still a considerable gap between academic developments and investment practice, and between normative and descriptive investment theories. In many ways, we have not yet experienced the unprecedented progress that took place in the 1980s and 1990s when academia and the derivatives industry challenged and worked by each other, leading to outstanding scientific progress in financial mathematics and quantitative finance.

The aim of this paper is to describe the main academic developments in portfolio management, discuss modeling issues, present various methods and expose some of the current challenges that the investment research faces.

2. Model certainty and investment management

Models of optimal investment management give rise to stochastic optimization problems with expected payoffs. There are three main ingredients in their specification: the model for the stochastic market environment, the investment horizon and the optimality criterion.

The market consists of assets whose prices are modelled as stochastic processes in an underlying probability space. The associated measure is known as the real, or historical, measure $\mathbb{P}$. Popular paradigms of prices are diffusion processes (2.2), (2.3), Itô processes (2.11) and, more generally, semimartingales (sections 3.1 and 3.2). When the price model is known we say that there is no model uncertainty.

The trading horizon is the time during which trading takes place, typically taken to have deterministic finite length. Depending on the application, the horizon can be infinitesimal (high frequency trading), short (hedge funds), medium (mutual funds) or long (pension funds). Models of infinite horizon have been also considered, especially when intermediate consumption is incorporated or when the criterion is asymptotic, like optimal long-term growth, risk-sensitive payoff and others.

The optimality criterion is built upon the utility function, a concept measuring risk and

uncertainty that dates back to D. Bernoulli (1738). He was the first to argue that utility should not be proportional to wealth but, rather, have decreasing marginal returns, thus, alluding for the first time to its concavity property. Bernoulli's pioneering ideas were rejected at that time and it took close to two centuries (with the exception of the work of Gossen) to be recognized. In 1936, Alt and few years later von Neumann and Morgenstern proposed the axiomatic foundation of expected utility and argued that the behavior of a rational investor must coincide with that of an individual who values random payoffs using an expected utility criterion. This normative work was further developed by Friedman and Savage, Pratt and Arrow. In the latter works, the quantification of individual aversion to risk - via the so called risk aversion coefficient - was proposed and few years later, Markowitz developed the influential "mean-variance" portfolio theory. In 1969, Merton built a continuous-time portfolio management model of expected utility for log-normal stock prices, and since then the academic literature in this area has seen substantial growth. We refer the reader to the review article [70] for further details and references.

The expected utility criterion enables us to quantify and rank the outcomes of investments policies π by mapping the wealth X_T^π they generate to its expected utility,

$$X_T^\pi \to E_{\mathbb{P}}\left(U\left(X_T^\pi\right)\right), \tag{2.1}$$

where $\mathbb{P}$ is the aforementioned historical measure and U a deterministic function that is smooth, strictly increasing and strictly concave, and satisfies appropriate asymptotic properties. The objective is then to *maximize* $E_{\mathbb{P}}\left(U\left(X_T^\pi\right)\right)$ over all admissible portfolios. The portfolios are the amounts (or proportions of current wealth) that are dynamically allocated to the different accounts. They are stochastic processes on their own and might satisfy (control) constraints, as it is discussed below.

There are two main directions in studying optimal portfolio problems. Under Markovian assumptions for the asset price processes, the value function is analyzed via PDE and stochastic control arguments applied to the associated Hamilton-Jacobi-Bellman (HJB) equation. We discuss this direction in detail next. For more general market settings, the powerful theory of duality is used. This approach yields elegant results for the value function and the optimal wealth. The optimal portfolios can be then characterized via martingale representation results for the optimal wealth process (see, among others, [27, 28, 30, 31, 55, 56]). We discuss the duality approach in sections 3.1 and 3.2 herein.

2.1. A diffusion market model and its classical (backward) expected utility criterion. We consider the popular paradigm in which trading takes place between a riskless asset (bond) and a risky one (stock). The stock price is modelled as a diffusion process whose coefficients depend on a correlated stochastic factor. Stochastic factors have been used in a number of academic papers to model the time-varying predictability of stock returns, the volatility of stocks as well as stochastic interest rates (for an extended bibliography, see the review article [67]).

From the technical point of view, a stochastic factor model is the simplest and most direct extension of the celebrated Merton model in which stock dynamics are taken to be log-normal (see [40]). However, as it is discussed herein, relatively little is known about the regularity of the value function, and the form and properties of the optimal policies once the log-normality assumption is relaxed and correlation between the stock and the factor is introduced. This is despite the Markovian nature of the problem at hand, the advances in the

theories of fully nonlinear PDE and stochastic optimization of controlled diffusion processes, as well as the available computational tools.

Specifically, complete results on the validity of the Dynamic Programming Principle, smoothness of the value function, existence and verification of optimal feedback controls, representation of the value function and numerical approximations are still lacking. The only cases that have been extensively analyzed are the ones of homothetic utilities (exponential, power and logarithmic). In these cases, convenient scaling properties reduce the HJB equation to a quasilinear one (even linear, see (2.9)). The analysis, then, simplifies considerably both from the analytic as well as the probabilistic points of view (see, for example, [52] and [66]).

The lack of rigorous results for the regularity and other properties of the value function, when the utility function is general, limits our understanding of the structure of the optimal policies. Informally speaking, the first-order conditions in the HJB equation yield that the optimal feedback portfolio consists of two components (see (2.7)). The first is the so-called *myopic portfolio* and has the same functional form as the one in the classical Merton problem. The second component, usually referred to as the *excess hedging demand*, is generated by the stochastic factor. Conceptually, very little is understood about this term. In addition, the sum of the two components may become zero which implies that it is optimal for a risk averse investor not to invest in a risky asset with positive risk premium. A satisfactory explanation for this counter intuitive phenomenon - related to the so-called market participation puzzle - is also lacking.

We continue with the description of the market model. The stock price S_t, $t \geq 0$, is modelled as a diffusion process solving

$$dS_t = \mu\left(Y_t\right) S_t dt + \sigma\left(Y_t\right) S_t dW_t^1, \tag{2.2}$$

with $S_0 > 0$. The stochastic factor Y_t, $t \geq 0$, satisfies

$$dY_t = b\left(Y_t\right) dt + d\left(Y_t\right)\left(\rho dW_t^1 + \sqrt{1-\rho^2} dW_t^2\right), \tag{2.3}$$

with $Y_0 = y$, $y \in \mathbb{R}$. The process $W_t = \left(W_t^1, W_t^2\right)$, $t \geq 0$, is a standard $2-$dim Brownian motion, defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The underlying filtration is $\mathcal{F}_t = \sigma\left(W_s : \ 0 \leq s \leq t\right)$, and it is assumed that $\rho \in (-1, 1)$. The market coefficients $f = \mu, \sigma, b$ and d satisfy global Lipschitz and linear growth conditions and the non-degeneracy condition $\sigma\left(y\right) \geq l > 0$, $y \in \mathbb{R}$. The riskless asset offers constant interest rate $r > 0$.

Starting with an initial endowment x, at time $t \in [0, T)$, the investor invests at future times $s \in (t, T]$ in the riskless and risky assets. The present value of the amounts allocated in the two accounts are denoted, respectively, by π_s^0 and π_s. The investor's (discounted) wealth is, then, given by $X_s^\pi = \pi_s^0 + \pi_s$. It follows that it satisfies $dX_s^\pi = \sigma\left(Y_s\right)\pi_s\left(\lambda\left(Y_s\right) ds + dW_s^1\right)$, where $\lambda\left(Y_s\right) = \frac{\mu(Y_s) - r}{\sigma(Y_s)}$.

A portfolio, π_s, is admissible if it is self-financing, $\mathcal{F}_s$-adapted, $E_{\mathbb{P}}\left(\int_t^T \sigma^2\left(Y_s\right)\pi_s^2 ds\right) < \infty$ and the associated discounted wealth satisfies the state constraint $X_s^\pi \geq 0, \mathbb{P}-$a.s. We denote the set of admissible strategies by $\mathcal{A}$.

Frequently, portfolio constraints are also present which further complicate the analysis. Notable cases are the so-called *drawdown* constraints, for which $X_t^\pi \geq \alpha \max_{0 \leq s \leq t} X_s^\pi$ with $\alpha \in (0, 1)$, *leverage* constraints, when $|\pi_t| \leq g\left(X_t^\pi\right)$ for an admissible function g, and *stochastic target* constraints, for which $X_T^\pi \geq Z_T$ for a random level Z_T.

The objective, known as the *value function* (or indirect utility), is formulated as

$$V(x,y,t;T) = \sup_{\mathcal{A}} E_{\mathbb{P}}\left(U\left(X_T^{\pi}\right)\middle|\, \mathcal{F}_t, X_t^{\pi} = x, Y_t = y\right), \tag{2.4}$$

for $(x,y,t) \in \mathbb{R}_+ \times \mathbb{R} \times [0,T]$. The utility function $U : \mathbb{R}_+ \to \mathbb{R}$ is $C^4(\mathbb{R}_+)$, strictly increasing and strictly concave, and satisfies certain asymptotic properties (see, among others, [55] and [56]). As solution of a stochastic optimization problem, the value function is expected to satisfy the Dynamic Programming Principle (DPP), namely,

$$V(x,y,t) = \sup_{\mathcal{A}} E_{\mathbb{P}}\left(V\left(X_s^{\pi}, Y_s, s\right)\middle|\, \mathcal{F}_t, X_t^{\pi} = x, Y_t = y\right), \tag{2.5}$$

for $s \in [t,T]$. This is a fundamental result in optimal control and has been proved for a wide class of optimization problems. For a detailed discussion on the validity (and strongest forms) of the DPP in problems with controlled diffusions, we refer the reader to [18] (see, also [6, 8, 14, 33, 35, 65]). Key issues are the measurability and continuity of the value function process as well as the compactness of the set of admissible controls. A weak version of the DPP was proposed in [9] where conditions related to measurable selection and boundness of controls are relaxed. Related results for the case of bounded payoffs can be found in [3] and, more recently, new results appeared in [71].

Besides its technical challenges, the DPP exhibits two important properties of the value function process. Specifically, the process $V\left(X_s^{\pi}, Y_s, s\right)$, $s \in [t,T]$, is a *supermartingale* for an arbitrary admissible investment strategy and becomes a *martingale* at an optimum (provided certain integrability conditions hold). Moreover, observe that the DPP yields a backward in time algorithm for the computation of the maximal expected utility, starting at expiration with U and using the martingality property to compute the solution conditionally for earlier times. For this, we occasionally refer to the classical problem as the *backward* one.

The Markovian assumptions on the stock price and stochastic factor dynamics allow us to study the value function via the associated HJB equation, stated in (2.6) below. Fundamental results in the theory of controlled diffusions yield that if the value function is smooth enough then it satisfies the HJB equation. Moreover, optimal policies may be constructed in a feedback form from the first-order conditions in the HJB equation, provided that the candidate feedback process is admissible and the wealth SDE has a strong solution when the candidate control is used. The latter usually requires further regularity on the value function. In the reverse direction, a smooth solution of the HJB equation that satisfies the appropriate terminal and boundary (or growth) conditions may be identified with the value function, provided the solution is unique in the appropriate sense. These results are usually known as the "verification theorem" and we refer the reader to [6, 8, 14, 33, 35, 65] for a general exposition on the subject.

In maximal expected utility problems, it is rarely the case that the arguments in either direction of the verification theorem can be established. Indeed, it is difficult to show a priori regularity of the value function, with the main difficulties coming from the lack of global Lipschitz regularity of the coefficients of the controlled process with respect to the controls and from the non-compactness of the set of admissible policies. It is, also, difficult to establish existence, uniqueness and regularity of the solutions to the HJB equation. This is caused primarily by the presence of the control policy in the volatility of the controlled wealth process which makes the classical assumptions of global Lipschitz conditions of the

equation with regards to the non linearities to fail. Additional difficulties come from state constraints and the non-compactness of the set of admissible portfolios.

Regularity results for the value function (2.4) for general utility functions have not been obtained to date except, as mentioned earlier, for the special cases of homothetic preferences. The most general result in this direction, and in a much more general market model, was obtained using duality methods in [32] where it is shown that the value function is twice differentiable in the spatial argument but without establishing the continuity of the derivative. Because of lack of general rigorous results, we proceed with an informal discussion about the optimal feedback policies. For the model at hand, the associated HJB equation is

$$V_t + \max_{\pi} \left(\frac{1}{2}\sigma^2 (y)\, \pi^2 V_{xx} + \pi \left(\mu (y)\, V_x + \rho\sigma (y)\, d (y)\, V_{xy} \right) \right) \tag{2.6}$$

$$+ \frac{1}{2} d^2 (y)\, V_{yy} + b (y)\, V_y = 0,$$

with $V (x, y, T) = U (x)$, $(x, y, t) \in \mathbb{R}_+ \times \mathbb{R} \times [0, T]$. The verification results would yield that under appropriate regularity and growth conditions, the feedback policy $\pi_s^* = \pi^* (X_s^*, Y_s, s)$, $s \in (t, T]$, with

$$\pi^* (x, y, t) = -\frac{\lambda (y)}{\sigma (y)} \frac{V_x (x, y, t)}{V_{xx} (x, y, t)} - \rho \frac{d (y)}{\sigma (y)} \frac{V_{xy} (x, y, t)}{V_{xx} (x, y, t)}, \tag{2.7}$$

and $X_s^{\pi^*}$ solving $dX_s^{\pi^*} = \sigma (Y_s)\, \pi \left(X_s^{\pi^*}, Y_s, s \right) \left(\lambda (Y_s)\, ds + dW_s^1 \right)$, is admissible and optimal.

Some answers to the questions related to the characterization of the solutions to the HJB equation may be given if one relaxes the requirement to have classical solutions. An appropriate class of weak solutions turns out to be the so called *viscosity solutions* ([11, 35, 36, 61]). Results related to the value function being the unique viscosity solution of (2.6) are rather limited. Recently, it was shown in [50] that the partial $V_x (x, y, t)$ is the unique viscosity solution of the marginal HJB equation. Other results, applicable for non-compact controls but for bounded payoffs, can be found in [3].

A key property of viscosity solutions is their robustness (see [36]). If the HJB has a unique viscosity solution (in the appropriate class), robustness is used to establish convergence of numerical schemes for the value function and the optimal feedback laws. Such numerical studies have been carried out successfully for a number of applications. However, for the model at hand, no such studies are available. Numerical results using Monte Carlo techniques have been obtained in [12] for a model more general than the one herein. More recently, the authors in [50] proposed a Trotter-Kato approximation scheme for the value function and an algorithm on how to construct $\varepsilon-$optimal portfolio policies.

Important questions arise on the dependence, sensitivity and robustness of the optimal feedback portfolio, especially of the excess hedging demand term, in terms of the market parameters, the wealth, the level of the stochastic factor and the risk preferences. Such questions are central in financial economics and have been studied, primarily in simpler models in which intermediate consumption is also incorporated. Recent results for more general models can be found, for example, in [34]. For diffusion models with a perfectly correlated stochastic factor, qualitative results can be found, among others, in [29] and [62] and for log-normal models in [7, 25, 42, 64]. However, a qualitative study for general utility functions and/or arbitrary factor dynamics has not been carried out to date. Another open

question, which is more closely related to applications, is how one could infer the investor's risk preferences from her investment targets. This is a difficult inverse problem and has been partially addressed in [41] and [45].

Example 2.1. A commonly used utility function is the homothetic $U(x) = \frac{x^\gamma}{\gamma}$, $x \geq 0$, $\gamma \in (0,1)$. In this case, the value function is given by (see [66])

$$V(x,y,t) = \frac{x^\gamma}{\gamma}\left(F(y,t)\right)^\delta \tag{2.8}$$

where $\delta = \frac{1-\gamma}{1-\gamma+\rho^2\gamma}$ and F solves the linear equation

$$F_t + \frac{1}{2}d^2(y)F_{yy} + \left(b(y) + \rho\frac{\gamma}{1-\gamma}\lambda(y)a(y)\right)F_y + \frac{1}{2}\frac{\gamma}{(1-\gamma)\delta}\lambda^2(y)F = 0, \tag{2.9}$$

with $F(y,T) = 1$. The Feynman-Kac formula then yields the probabilistic representation

$$V(x,y,t) = \frac{x^\gamma}{\gamma}\left(E_{\bar{\mathbb{P}}}\left(e^{\int_t^T \frac{1}{2}\frac{\gamma}{(1-\gamma)\delta}\lambda^2\left(\bar{Y}_s\right)ds}\middle|\, \bar{Y}_t = y\right)\right)^\delta \tag{2.10}$$

where $\bar{Y}_t$, $t \in [0,T]$, solves $d\bar{Y}_t = (b\left(\bar{Y}_t\right) + \rho\frac{\gamma}{1-\gamma}\lambda\left(\bar{Y}_t\right)a\left(\bar{Y}_t\right))dt + d\left(\bar{Y}_t\right)dW_t^{\bar{\mathbb{P}}}$, with $W^{\bar{\mathbb{P}}}$ being a standard Brownian motion under a measure $\bar{\mathbb{P}}$.

2.2. An Itô market model and its forward performance criterion. Besides the difficulties discussed earlier, there are other issues that limit the development of a flexible enough optimal investment theory in complex market environments. One of them is the "static" choice of the utility function at the specific investment horizon. Indeed, once the utility function is a priori specified, no revision of risk preferences is possible at any intermediate trading time. In addition, once the horizon is chosen, no investment performance criteria can be formulated for horizons longer than the initial one. As a result, extending the investment horizon (due to new incoming investment opportunities, change of risk attitude, unpredictable price shocks, etc.) is not possible.

Addressing these limitations has been the subject of a number of studies and various approaches have been proposed. With regards to the horizon length, the most popular alternative has been the formulation of the investment problem in $[0,\infty)$ and either incorporating intermediate consumption or optimizing the investor's long-term optimal behavior. Investment modes with random horizon have been also considered, and the revision of risk preferences has been partially addressed by recursive utilities (see, for example, [13] and [59]).

An alternative approach which addresses both shortcomings of the expected utility approach has been proposed recently by the author and Musiela (see, [43–45]). The associated criterion, the so called *forward performance process,* is developed in terms of a family of utility fields defined on $[0,\infty)$ and indexed by the wealth argument. Its key properties are the (local) martingality at an optimum and (local) supermartingality away from it. These are in accordance with the analogous properties of the classical value function process, we discussed earlier, which stem out from the Dynamic Programming Principle (cf. (2.5)). Intuitively, the average value of an optimal strategy at any future date, conditional on today's information, preserves the performance of this strategy up until today. Any strategy that fails to maintain the average performance over time is, then, sub-optimal. We refer the reader to

[44] and [45] for further discussion on this new concept and its connection with the classical expected utility theory.

Next, we introduce the definition of the forward performance process and present old and more recent results. The market environment consists of one riskless security and k stocks. For $i = 1, ..., k$, the stock price S_t^i, $t > 0$, is an Itô process solving

$$dS_t^i = S_t^i \left(\mu_t^i dt + \sigma_t^i \cdot dW_t^j \right) \tag{2.11}$$

with $S_0^i > 0$. The process $W_t = \left(W_t^1, ..., W_t^k\right)$ is a standard $d-$dim Brownian motion, defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathcal{F}_t = \sigma\left(W_s : 0 \leq s \leq t\right)$. The coefficients μ_t^i and σ_t^i, $i = 1, ..., k$, are $\mathcal{F}_t-$adapted processes with values in $\mathbb{R}$ and $\mathbb{R}^d$, respectively. For brevity, we denote by σ_t the volatility matrix, i.e., the $d \times k$ random matrix $\left(\sigma_t^{ji}\right)$, whose i^{th} column represents the volatility σ_t^i of the i^{th} risky asset. The riskless asset has the price process B satisfying $dB_t = r_t B_t dt$ with $B_0 = 1$, and a nonnegative $\mathcal{F}_t-$adapted interest rate process r_t. Also, we denote by μ_t the $k \times 1$ vector with coordinates μ_t^i. The processes μ_t, σ_t and r_t satisfy the appropriate integrability conditions and it is further assumed that $(\mu_t - r_t \mathbf{1}) \in Lin\left(\sigma_t^T\right)$.

The market price of risk is given by the vector $\lambda_t = \left(\sigma_t^T\right)^+ (\mu_t - r_t \mathbf{1})$, where $\left(\sigma_t^T\right)^+$ is the Moore-Penrose pseudo-inverse of σ_t^T. It is assumed that, for all $t > 0$, $E_{\mathbb{P}} \int_0^t |\lambda_s|^2 \, ds < \infty$.

Starting at $t = 0$ with an initial endowment $x \in \mathbb{D}$, $\mathbb{D} \subseteq [-\infty, \infty]$, the investor invests dynamically among the assets. The (discounted) value of the amounts invested are denoted by π_t^0 and π_t^i, $i = 1, ..., k$, respectively. The (discounted) wealth process is, then, given by $X_t^\pi = \sum_{i=0}^k \pi_t^i$, and satisfies

$$dX_t^\pi = \sum_{i=1}^k \pi_t^i \sigma_t^i \cdot (\lambda_t dt + dW_t) = \sigma_t \pi_t \cdot (\lambda_t dt + dW_t), \tag{2.12}$$

where the (column) vector, $\pi_t = \left(\pi_t^i;\ i = 1, ..., k\right)$. The admissibility set, $\mathcal{A}$, consists of self-financing $\mathcal{F}_t-$adapted processes π_t such that $E_{\mathbb{P}} \int_0^t |\sigma_s \pi_s|^2 \, ds < \infty$, and $X_t^\pi \in \mathbb{D}$, for $t \geq 0$.

The initial datum is taken to be a strictly concave and strictly increasing function of wealth, $u_0 : \mathbb{D} \to \mathbb{R}$ with $u_0 \in \mathcal{C}^4(\mathbb{D})$. The specification of admissible initial conditions deserves special attention and is discussed later (see (2.20)).

Next, we present the definition of the forward performance process. The one below is a relaxed version of the original definition, given in [44]**,** where stronger integrability conditions were needed.

Definition 2.2. An $\mathcal{F}_t-$adapted process $U(x,t)$ is a local forward performance process if for $t \geq 0$ and $x \in \mathbb{D}$:

i) the mapping $x \to U(x,t)$ is strictly concave and strictly increasing,

ii) for each $\pi \in \mathcal{A}$, the process $U\left(X_t^\pi, t\right)$ is a local supermartingale, and

iii) there exists $\pi^* \in \mathcal{A}$ such that the process $U\left(X_t^{\pi^*}, t\right)$ is a local martingale.

Variations of the above definition have appeared, among others, in [15] and [49]. In [69], the alternative terminology "*self-generating*"was introduced, for the forward performance

satisfies, for *all* $0 \leq t \leq s$,

$$U(x,t) = ess\sup_{\mathcal{A}} E_{\mathbb{P}}\left(U\left(X_s^{\pi}, s\right)\middle|\, \mathcal{F}_t,\ X_t^{\pi} = x\right). \tag{2.13}$$

Note that in the classical (backward) case $(0 \leq t \leq s \leq T)$ the above property is a direct *consequence* of the DPP. In the forward framework, however, it *defines* the forward performance process. Clearly, if for the backward problem with finite horizon T one uses as terminal utility $U_T(x) = U(x,T)$, the backward and the forward problems coincide on $[0,T]$.

The axiomatic construction of forward performance is an *open* problem, and results have been derived only for the exponential case (see [69]). More recently, the authors in [49] proposed a class of forward performances processes that are deterministic functions of underlying stochastic factors (see, for example, (2.24) herein).

2.2.1. Stochastic PDE for the forward performance process. In [46] a stochastic PDE was derived as a sufficient condition for a process to be a forward performance. In many aspects, the forward SPDE is the analogue of the HJB equation that appears in the classical theory of stochastic optimization.

Proposition 2.3.

i) *Let $U(x,t)$, $(x,t) \in \mathbb{D}\times[0,\infty)$, be an $\mathcal{F}_t-$adapted process such that the mapping $x \to U(x,t)$ is strictly concave, strictly increasing and smooth enough so that the Itô-Ventzell formula can be applied to $U\left(X_t^{\pi},t\right)$, for any strategy $\pi \in \mathcal{A}$. Let us, also, assume that the process $U(x,t)$ satisfies*

$$dU(x,t) = \frac{1}{2}\frac{\left|U_x(x,t)\lambda_t + \sigma_t\sigma_t^{+}a_x(x,t)\right|^2}{U_{xx}(x,t)}dt + a(x,t)\cdot dW_t, \tag{2.14}$$

where the volatility $a(x,t)$ is an $\mathcal{F}_t-$adapted, $d-$dimensional and continuously differentiable in the spatial argument process. Then, $U\left(X_t^{\pi},t\right)$ is a local supermartingale for every admissible portfolio strategy π.

ii) *Assume that the stochastic differential equation*

$$dX_t = -\frac{U_x(X_t,t)\lambda_t + \sigma_t\sigma_t^{+}a_x(X_t,t)}{U_{xx}(X_t,t)}\cdot\left(\lambda_t dt + dW_t\right)$$

has a solution X_t, with $X_0 = x$, and $X_t \in \mathbb{D}$, $t \geq 0$, and that the strategy π_t^*, $t \geq 0$, defined by

$$\pi_t^* = -\sigma_t^{+}\frac{U_x(X_t,t)\lambda_t + a_x(X_t,t)}{U_{xx}(X_t,t)}$$

is admissible. Then, X_t corresponds to the wealth generated by this investment strategy, that is $X_t = X_t^{\pi^*}$, $t > 0$. The process $U\left(X_t^{\pi^*},t\right)$ is a local martingale and, hence, $U(x,t)$ is a local forward performance value process. The process π_t^* is optimal.

An important ingredient of the forward SPDE is the *forward volatility process* $a(x,t)$. This is a novel model input that is up to the investor to choose, in contrast to the classical

value function process whose volatility process is uniquely determined from its Itô decomposition. In general, the forward volatility may depend explicitly on t, x, U and its derivatives, as it is, for instance, shown in the examples below. More general dependencies and admissible volatility representations have been proposed in [15].

The initial condition $u_0(x)$ is an additional model input. In contrast to the classical framework where the class of admissible (terminal) utilities is rather large, the family of admissible forward initial data can be rather restricted.

The analysis of the forward performance SPDE (2.14) is a formidable task. The reasons are threefold. Firstly, it is not only degenerate and fully nonlinear but is, also, formulated forward in time, which might lead to "ill-posed" behavior. Secondly, one needs to specify the appropriate class of admissible volatility processes, namely, volatility inputs that generate strictly concave and strictly increasing solutions of (2.14). The volatility specification is quite difficult both from the modelling and the technical points of view. Thirdly, as mentioned earlier, one also needs to specify the appropriate class of initial conditions $u_0(x)$. As it has been shown in [45] and discussed in the sequel, even the simple case of zero volatility poses a number of challenges.

Addressing these issues is an ongoing research effort of several authors; see, among others, in [4, 15, 16, 46, 49] and [51].

2.2.2. The time-monotone case and its variants. A fundamental class of forward performance processes are the ones that correspond to non-volatile criteria, $a(x,t) \equiv 0$, $t \geq 0$. The forward performance SPDE (2.14) simplifies to

$$dU(x,t) = \frac{1}{2}|\lambda_t|^2 \frac{U_x^2(x,t)}{U_{xx}(x,t)}dt, \tag{2.15}$$

and, thus, its solutions are processes of finite variation. In particular, they are decreasing in time, as it follows from the strict concavity requirement. The analysis of these processes was carried out in [45], and we highlight the main results next.

There are three functions that play pivotal role in the construction of the forward performance process, as well as of the optimal wealth and optimal portfolio processes. The first function is $u : \mathbb{D} \times [0,\infty) \to \mathbb{R}$, with $u \in C^{4,1}(\mathbb{D} \times [0,\infty))$, solving the HJB type equation

$$u_t = \frac{1}{2}\frac{u_x^2}{u_{xx}}, \tag{2.16}$$

and satisfying an admissible initial condition, $U(x,0) = u_0(x)$ (see (2.20)).

The second function is the so-called local absolute risk tolerance $r : \mathbb{D} \times [0,\infty) \to \mathbb{R}_+$, defined by $r(x,t) = -\frac{u_x(x,t)}{u_{xx}(x,t)}$. It solves an autonomous fast-diffusion type equation, $r_t + \frac{1}{2}r^2 r_{xx} = 0$, with $r(x,0) = -\frac{u_0'(x)}{u_0''(x,t)}$.

The third is an increasing space-time harmonic function, $h : \mathbb{R} \times [0,\infty) \to \mathbb{D}$, defined via a Legendre-Fenchel type transformation

$$u_x(h(x,t),t) = e^{-x+\frac{1}{2}t}. \tag{2.17}$$

It solves the (backward) heat equation

$$h_t + \frac{1}{2}h_{xx} = 0, \tag{2.18}$$

with initial condition $h(x,0) = \left(u_0'\right)^{(-1)}(e^{-x})$.

Using the classical results of Widder (see [63]) for the representation of positive solutions[1] of (2.18), it follows that $h(x,t)$ must be given in the integral form

$$h(x,t) = \int_S \frac{e^{yx-\frac{1}{2}y^2t}-1}{y}\nu(dy), \tag{2.19}$$

where ν is a positive, finite, Borel measure with support $S \in [-\infty,\infty]$. Detailed analysis on the interplay among the support S, the range of h, the structure and the asymptotic properties of u can be found in [45]. It was also shown therein that there is a one-to-one correspondence between such solutions of (2.18) to strictly increasing and strictly concave solutions of (2.16) (see, Propositions 9, 13 and 14**).**

One then sees that the measure ν becomes the defining element in the entire construction, for it determines the function h and, in turn, u and r. How this measure could be extracted from various distributional investment targets is an interesting question and has been discussed in [41] and [45].

We also see that the definition (cf. (2.17)) of the auxiliary function h and its structural representation (2.19) dictate that the initial utility $u_0(x)$, $x \in \mathbb{D}$, is given by

$$\left(u_0'\right)^{(-1)}(x) = \int_S \frac{e^{-y\ln x}-1}{y}\nu(dy). \tag{2.20}$$

In other words, only utilities whose inverse marginals have the above form can serve as initial conditions. Characterizing the set of *admissible initial data* $u_0(x)$ for general volatile performance criteria and, moreover, provide an intuitively meaningful financial interpretation for them is an interesting open question.

We summarize the general results next. As (2.21) and (2.22) below show, one obtains rather *explicit* stochastic representations of the optimal wealth and portfolio policies, despite the ill-posedness of the underlying problem, the complexity of the price dynamics, and the path-dependence nature of all quantities involved.

Proposition 2.4. *Let $u : \mathbb{D}\times[0,\infty) \to \mathbb{R}$ be a strictly increasing and strictly concave solution of (2.16), satisfying an admissible initial condition $u(x,0) = u_0(x)$, and $r(x,t)$ be its local absolute risk tolerance function. Let also $h : \mathbb{R}\times[0,\infty) \to \mathbb{D}$ be the associated harmonic function (cf. (2.17)). Define the market-input processes A_t and M_t, $t \geq 0$, as*

$$M_t = \int_0^t \lambda_s \cdot dW_s \quad \text{and} \quad A_t = \langle M\rangle_t = \int_0^t |\lambda_s|^2\, ds.$$

Then, the process $U(x,t) = u(x,A_t)$, $t \geq 0$, is a forward performance. Moreover, the optimal portfolio process is given by

$$\pi_t^{*,x} = r\left(X_t^{\pi^*}, A_t\right)\sigma_t^+\lambda_t = h_x\left(h^{(-1)}(x,0) + A_t + M_t, A_t\right)\sigma^+\lambda. \tag{2.21}$$

The optimal wealth process $X_t^{\pi^*}$ solves $dX_t^{\pi^*} = \sigma_t\sigma_t^+\lambda_t r\left(X_t^{\pi^*}, A_t\right)\cdot(\lambda_t dt + dW_t)$ with $X_0^{\pi^*} = x$, and is given by

$$X_t^{\pi^*} = h\left(h^{(-1)}(x,0) + A_t + M_t, A_t\right). \tag{2.22}$$

[1]Widder's results are not applied to $h(x,t)$ directly, for it might not be positive, but to its space derivative $h_x(x,t)$.

Representations (2.21) and (2.22) enable us to study the optimal processes in more detail. Among others, one can draw analogies between option prices and their sensitivities (gamma, delta and other "greeks") and study analogous quantities for the optimal investments. Moreover, one can study the distribution of hitting times of the optimal wealth, calculate its moments, running maximum, Value at Risk, expected shortfall and other investment performance markers.

Example 2.5.

i) Let $\mathbb{D} = \mathbb{R}$ and $\nu = \delta_0$, where δ_0 is a Dirac measure at 0. Then, $h(x,t) = x$ and $u(x,t) = 1 - e^{-x+\frac{t}{2}}$. The forward performance process is, for $t \geq 0$, $U(x,t) = 1 - e^{-x+\frac{A_t}{2}}$ (see [43] and [69]).

ii) Let $\mathbb{D} = \mathbb{R}_+$ and $\nu = \delta_\gamma$, $\gamma > 1$. Then $h(x,t) = \frac{1}{\gamma} e^{\gamma x - \frac{1}{2}\gamma^2 t}$. Since $\nu((0,1]) = 0$, it turns out that $u(x,t) = kx^{\frac{\gamma-1}{\gamma}} e^{-\frac{\gamma-1}{2}t}$, $k = \frac{1}{\gamma-1}\gamma^{\frac{\gamma-1}{\gamma}}$. The forward performance process is, for $t \geq 0$,

$$U(x,t) = kx^{\frac{\gamma-1}{\gamma}} e^{-\frac{\gamma-1}{2}A_t}. \tag{2.23}$$

There exist two interesting *variants* of the time-monotone forward performance process, which correspond to *non-zero* volatility processes. To this end, consider the auxiliary processes Y_t, Z_t, $t \geq 0$, solving

$$dY_t = Y_t \delta_t \cdot (\lambda_t dt + dW_t) \quad \text{and} \quad dZ_t = Z_t \varphi_t \cdot dW_t,$$

with $Y_0 = Z_0 = 1$ and the coefficients δ_t and φ_t being $\mathcal{F}_t-$adapted and bounded (by a deterministic constant) processes. We further assume that $\delta_t, \varphi_t \in Lin(\sigma_t)$.

- *The benchmark case:* $a(x,t) = -xU_x(x,t)\delta_t$. Then, $U(x,t) = u\left(\frac{x}{Y_t}, A_t^{(\delta)}\right)$ with $A_t^{(\delta)} = \int_0^t |\lambda_s - \delta_s|^2 ds$ is a forward performance process. The factor Y_t normalizes the wealth argument and, thus, can be thought as a *benchmark* (or a numeraire) in relation to which one might wish to measure the performance of investment strategies.
- *The market-view case*: $a(x,t) = U(x,t)\varphi_t$. Then, $U(x,t) = u\left(x, A_t^{(\varphi)}\right) Z_t$ with $A_t^{(\varphi)} = \int_0^t |\lambda_s + \varphi_s|^2 ds$ is a forward performance process. The factor Z_t can be thought as a device offering flexibility to the forward solutions in terms of the asset returns. This might be needed if the investor has different *views* about the future market movements or faces trading constraints. In such cases, the returns need to be modified which essentially points to a change of measure, away from the historical one. This is naturally done through an exponential martingale.

2.2.3. The stochastic factor case and its forward volatility process. We now revert to the stochastic factor example with dynamics (2.2) and (2.3), studied earlier under the classical (backward) formulation, and we examine its forward analogue. To this end, consider a process $U(x,t)$, $t \geq 0$, given by

$$U(x,t) = v(x, Y_t, t), \tag{2.24}$$

for a deterministic function $v : \mathbb{R}_+ \times \mathbb{R} \times [0,\infty)$. Then, the SPDE (2.14) takes the form

$$dU(x,t) = \frac{1}{2} \frac{\left(\lambda(Y_t) v_x(x, Y_t, t) + \rho d(Y_t) v_{xy}(x, Y_t, t)\right)^2}{v_{xx}(x, Y_t, t)} dt$$

$$+\rho d\left(Y_t\right) v_y\left(x, Y_t, t\right) dW_t^1 + \sqrt{1-\rho^2} d\left(Y_t\right) v_y\left(x, Y_t, t\right) dW_t^2,$$

with the forward volatility given by $a\left(x,t\right) = \left(\rho, \sqrt{1-\rho^2}\right) d\left(Y_t\right) v_y\left(x, Y_t, t\right)$. One then sees that if v satisfies (2.6) but now with an admissible initial (and not terminal) condition, say $v\left(x,y,0\right) = u_0\left(x\right)$, the process given in (2.24) is a forward performance. Solving (2.6) with an initial condition is an open problem because it not only inherits the difficulties discussed in the previous section but, now, one needs to deal with the ill-posedness of the HJB equation.

The homothetic case $u_0\left(x\right) = \frac{x^\gamma}{\gamma}$, $\gamma \in \left(0,1\right)$, has been extensively studied in [51]. Therein, it is shown that the forward performance process is given by an analogous to (2.8) formula, namely,

$$U\left(x,t\right) = \frac{1}{\gamma} x^\gamma \left(f\left(Y_t, t\right)\right)^\delta \tag{2.25}$$

provided that $f\left(y,t\right)$ satisfies the linear equation (2.9) with initial (and not terminal) condition $f\left(x,0\right) = 1$. This problem is more general than (2.18) due to the form of its coefficients, and, thus, more involved arguments needed to be developed. The multi-dimensional analogue of (2.25) was recently analyzed in [49]. Therein, $f\left(y,t\right)$ solves a multi-dimensional ill-posed linear problem with state-dependent coefficients. For such problems, there is no standard existence theory. The authors addressed this by developing a generalized version of the classical Widder's theorem.

Forward versus backward homothetic utilities. It is worth commenting on the different features of the three homothetic performance processes (2.10), (2.23) and (2.25). Tthe traditional value function (2.10) requires, for each $s \in \left[t,T\right)$ *forecasting* of the market price of risk in the *remaining trading horizon* $\left[s,T\right)$. In contrast, both (2.23) and (2.25) are constructed *path-by-path, given the information for the market price of risk up to today, in* $\left[0,s\right]$. The process (2.23) is decreasing in time, while (2.25) is not.

3. Model uncertainty and investment management

In the previous section, a prevailing assumption was that the historical measure $\mathbb{P}$ is a priori known. This, however, has been challenged by a number of scholars and gradually led to the development of selection criteria under *model uncertainty,* otherwise known as *ambiguity* or *Knightian uncertainty.* Pathbreaking work was done by Gilboa and Schmeidler in [22] and [58] who built an axiomatic approach for preferences not only towards risk - as it was done by von Neumann and Morgenstern for (2.1) - but also towards model ambiguity. They argued that such preferences can be numerically represented by a *"coherent" robust* utility functional of the form

$$X_T^\pi \to \inf_{Q \in \mathcal{Q}} E_Q\left(U\left(X_T^\pi\right)\right), \tag{3.1}$$

where U is a classical utility function and $\mathcal{Q}$ a family of probability measures. These measures can be thought as corresponding to different "prior" market models and the above infimum serves as the "worst-case scenario" in model misspecification.

A standard criticism for the above criterion, however, is that it allows for very limited, if at all, differentiation of models with respect to their plausibility. As discussed in [57], if, for instance, the family of prior models is generated from a confidence set in statistical

estimation, models with higher plausibility must receive a higher weight than models in the boundary of the confidence set. Furthermore, one should be able to incorporate information from certain stress test models and observed discrepancies with outcomes of models of possible priors. Such shortcomings of criterion (3.1) stem primarily from the axiom of certainty independence in [22]. Maccheroni et al. [37] relaxed this axiom and proposed a numerical representation of the form

$$X_T^\pi \to \inf_{Q\in\mathcal{Q}} \left(E_Q\left(U\left(X_T^\pi\right)\right)+\gamma\left(Q\right)\right), \tag{3.2}$$

where U is a classical utility function and the functional $\gamma\left(Q\right)$ serves as a *penalization* weight to each Q-market model.

The specification and representation of robust preferences and their penalty functionals have recently attracted considerable attention. It turns out that there is a deep connection between them, monetary utility functionals and risk measures. The latter, denoted by $\varphi\left(X\right)$ and $\rho\left(X\right)$, respectively, are mappings on financial positions X, represented as random variables on a given probability space $(\Omega,\mathcal{F},\mathbb{P})$ with $X\in L^\infty$. They are related as $\varphi\left(X\right)=-\rho\left(X\right)$.

Coherent risk measures were first introduced in [1] and were later extended to their convex analogues by [19, 21, 23]. Risk measures constitute one of the most active areas in financial mathematics with a substantial volume of results involving several areas in mathematics spanning from capacity theory and Choquet integration to BSDE, nonlinear expectations and stochastic differential games.

The (minimal) penalty function associated with a convex risk measure and its associated concave monetary utility functional, is defined, for probability measures $Q\ll\mathbb{P}$, by

$$\gamma\left(Q\right)=\sup_{X\in L^\infty}\left(E_Q\left(-X\right)-\rho\left(X\right)\right)=\sup_{X\in L^\infty}\left(\varphi\left(X\right)-E_Q\left(X\right)\right). \tag{3.3}$$

Extending criterion (3.1) to (3.2) is in direct analogy to generalizing the coherent risk measures to their convex counterparts. There is a substantial body of work on representation results for (3.3) which is, however, beyond the scope of this article.

Recent generalizations to (3.2) include the case

$$X_T^\pi \longrightarrow \inf_{Q\ll\mathbb{P}} G\left(Q,E_Q\left(U\left(X_T^\pi\right)\right)\right), \tag{3.4}$$

where G is the dual function in the robust representation of a *quasi-concave* utility functional.

In the sequel, we provide representative results on portfolio selection under the classical robust criterion (3.2) and its recently developed robust forward analogue.

3.1. Classical robust portfolio selection. The problem of portfolio selection in a finite horizon $[0,T]$ with the coherent robust utility (3.1) was studied by [53], [60] and others. Its generalization, corresponding to the convex analogue (3.2), was analyzed, among others, in [57] and we present below some of the results therein.

For an extensive overview of robust preferences and robust portfolio choice we refer the reader to the review paper [20].

The market model in [57] is similar to the standard semimartingale model in [30] and [31]. There is one riskless and d risky assets available for trading in $[0,T]$, $T<\infty$. The discounted price processes are modelled by a $d-$dim semimartingale $S_t=\left(S_t^1,...,S_t^d\right)$,

$t \in [0,T]$, on a filtered probability space $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P}\right)$. For $t \in [0,T]$, the control policies $\alpha_t = \left(\alpha_t^1, ..., \alpha_t^d\right)$ are self-financing, predictable and $S-$integrable processes. The associated discounted wealth process, X_t^α, is then given by $X_t^\alpha = x + \int_0^t \alpha_s \cdot dS_s$, and needs to satisfy $X_t^\alpha \geq 0$, $t \in [0,T]$. This formulation is slightly different than the ones in sections 2.1 and 2.2 in that the controls α_t now denote the number of shares (and not the discounted amounts) held at time t in the stock accounts.

For $x > 0$, $\mathcal{X}(x)$ stands for the set of all discounted wealth processes satisfying $X_0 \leq x$ and $X_t \geq 0$, $t \in (0,T]$. The classical (absence of arbitrage) model assumption is that $\mathcal{M} \neq \varnothing$, where $\mathcal{M}$ denotes the measures equivalent to $\mathbb{P}$ under which each $X_t \in \mathcal{X}(1)$, $t \in (0,T]$, is a local martingale (see [30]).

The value function of the *robust portfolio selection* problem is then defined, for $x \geq 0$, as

$$v(x) = \sup_{X\in\mathcal{X}(x)} \inf_{Q\in\mathcal{Q}} \left(E_Q\left(U\left(X_T\right)\right) + \gamma\left(Q\right)\right), \tag{3.5}$$

where γ is a minimal penalty function as in (3.3) and $\mathcal{Q} = \{Q \ll \mathbb{P} | \, \gamma(Q) < \infty\}$.

Because of the semimartingale assumption for the stock prices, classical stochastic optimization arguments do not apply and the *duality approach* comes in full force. As mentioned in the previous section, this approach has been extensively applied to portfolio choice problems and provides general characterization results of the value function and optimal policies through the dual problem, which is in general easier to analyze. There is a rich body of work in this area and we refer the reader, among others, to the classical references [28, 30, 31].

In the presence of model ambiguity, there is an extra advantage in using the duality approach because the dual problem simply involves the minimization of a convex functional while the primal one requires to find a saddle point of a functional which is concave in one argument and convex in the other.

We now describe the main notions and results in [57]. We stress, however, that for the ease of presentation we abstract from a number of detailed modeling assumptions and technical conditions.

We recall that the convex conjugate of the utility function U is defined, for $y > 0$, as $\tilde{U}(y) = \sup_{x>0}\left(U(x) - xy\right)$. Then, for every measure Q, $u_Q(x) = \sup_{X\in\mathcal{X}(x)} E_Q\left(U(X_T)\right)$ is a traditional value function as in (2.4). It was established in [30] that, for $Q \sim \mathbb{P}$ with finite primal value function $u_Q(x)$, the bidual relationships $u_Q(x) = \inf_{y>0}\left(\tilde{u}_Q(y) + xy\right)$ and $\tilde{u}_Q(y) = \sup_{x>0}\left(u_Q(x) - xy\right)$ hold, where the *dual value function* $\tilde{u}_Q(y)$ is given by $\tilde{u}_Q(y) = \inf_{Y\in\mathcal{Y}_Q(y)} E_Q\left(\tilde{U}\left(Y_T\right)\right)$, for $Q \in \mathcal{Q}$. The space $\mathcal{Y}_Q(y)$ is the set of all positive $Q-$supermartingales such that $Y_0 = y$ and the product XY is a $Q-$supermartingale for all $X \in \mathcal{X}(1)$.

In analogy, one then defines in [57] the *dual function of the robust portfolio* problem by

$$\tilde{u}(y) = \inf_{Q\in\mathcal{Q}}\left(\tilde{u}_Q(y) + \gamma(Q)\right) = \inf_{Q\in\mathcal{Q}} \inf_{Y\in\mathcal{Y}_Q(y)} \left(E_Q\left(\tilde{U}\left(Y_T\right)\right) + \gamma(Q)\right).$$

Then, for $y > 0$ such that $\tilde{u}(y) < \infty$, a pair (Q, Y) is a solution to the dual convex robust problem if $Q \in \mathcal{Q}$, $Y \in \mathcal{Y}_Q(y)$ and $\tilde{u}(y) = E_Q(\tilde{U}(Y_T)) + \gamma(Q)$. Let also $\mathcal{Q}^e = \{Q \in \mathcal{Q} | \, Q \sim \mathbb{P}\}$.

Theorems 2.4 and 2.6 in [57] provide characterization results for the primal and dual robust value functions, as well as for the optimal policies. In the next two propositions, we

highlight some of their main results.

Proposition 3.1. *Assume that for some $x > 0$ and $Q_0 \in \mathcal{Q}^e$, $u_{Q_0}(x) < \infty$ and that $\tilde{u}(y) < \infty$ implies that, for some $Q_1 \in \mathcal{Q}^e$, $\tilde{u}_{Q_1}(y) < \infty$. Then, the robust value function $u(x)$ is concave and finite, and satisfies*

$$u(x) = \sup_{X \in \mathcal{X}(x)} \inf_{Q \in \mathcal{Q}} \left(E_Q \left(U \left(X_T \right) \right) + \gamma \left(Q \right) \right) = \inf_{Q \in \mathcal{Q}} \sup_{X \in \mathcal{X}(x)} \left(E_Q \left(U \left(X_T \right) \right) + \gamma \left(Q \right) \right).$$

Moreover, the primal and the dual robust value functions u and $\tilde{u}$ satisfy

$$u(x) = \inf_{y>0} \left(\tilde{u}(y) + xy \right) \quad \text{and} \quad \tilde{u}(y) = \sup_{x>0} \left(u(x) - xy \right).$$

If $\tilde{u}(y) < \infty$, then the dual problem admits a solution, say (Q^*, Y^*) that is maximal, in that any other solution (Q, Y) satisfies $Q \ll Q^*$ and $Y_T = Y_T^*$, $Q-$a.s.

Note that the optimal measure Q^* might not be equivalent to $\mathbb{P}$ (see, for instance, example 3.2 in [57]). In such cases, one can show that the Q^*-market may admit arbitrage opportunities.

The existence of optimal policies requires the additional assumption that for all $y > 0$ and each $Q \in \mathcal{Q}^e$ the dual robust value function satisfies $\tilde{u}_Q(y) < \infty$.

Proposition 3.2. *For any $x > 0$, there exists an optimal strategy $X^* \in \mathcal{X}(x)$ for the robust portfolio selection problem. If $y > 0$ is such that $\tilde{u}'(y) = -x$ and (Q^*, Y^*) is a solution of the dual problem, then $X_T^* = I\left(Y_T^*\right)$, Q^*-a.s. for $I(x) = -\tilde{U}'(x)$, and (Q^*, Y^*) is a saddle point for the primal robust problem,*

$$u(x) = \inf_{Q \in \mathcal{Q}} \left(E_Q \left(U \left(X_T^* \right) \right) + \gamma \left(Q \right) \right) = E_{Q^*} \left(U \left(X_T^* \right) \right) + \gamma \left(Q^* \right) = u_{Q^*}(x) + \gamma \left(Q^* \right).$$

Furthermore, the product $X_t^* Y_t^* Z_t^*$ is a martingale under $\mathbb{P}$, where Z_t^*, $t \in [0,T]$, is the density process of Q^* with respect to $\mathbb{P}$.

Example 3.3. Examples of penalty functionals

- *Coherent penalties:* γ takes the values 0 or ∞. Then, (3.2) reduces to (3.1).
- *Entropic penalties:* $\gamma(Q) = H(Q|\mathbb{P})$, where the entropy function H is defined, for $Q \ll \mathbb{P}$, as

$$H(Q|\mathbb{P}) = \int \frac{dQ}{d\mathbb{P}} \ln \left(\frac{dQ}{d\mathbb{P}} \right) d\mathbb{P} = \sup_{Y \in L^\infty} \left(E_Q(Y) - \ln E_{\mathbb{P}} \left(e^Y \right) \right). \tag{3.6}$$

  In this case, $\inf_{Q \in \mathcal{Q}} \left(E_Q \left(U \left(X_T \right) \right) + \gamma \left(Q \right) \right) = \ln E_{\mathbb{P}} \left(e^{-U(X_T)} \right)$ and the robust portfolio problem (3.5) reduces to the standard one of maximizing $E_{\mathbb{P}} \left(e^{-U(X_T)} \right)$.

- *Dynamically consistent penalties:* $\gamma_t(Q) = E_Q \left(\int_t^T h(\eta_s)\, ds \middle| \mathcal{F}_t \right)$, $t \in [0,T)$, where the filtration $(\mathcal{F}_t)_{t \in [0,T]}$ is generated by a $d-$dim Brownian motion. Then, for every measure $Q \ll \mathbb{P}$, there exists a d-dim predictable process η_t with $\int_0^T |\eta_t|^2 \, dt < \infty$, Q-a.s. and $\frac{dQ}{d\mathbb{P}} = \mathcal{E}\left(\int_0 \eta_t \cdot dW_t \right)_T$ where $\mathcal{E}(M)_t = \exp\left(M_t - \langle M \rangle_t \right)$ for a continuous semimartingale M_t. The function h satisfies appropriate regularity and growth conditions (see example 3.4 in [57]). The specific choice $h(x) = \frac{1}{2}|x|^2$ corresponds to (3.6).

- *Shortfall risk penalties:* $\gamma(Q) = \inf_{\lambda>0}\left(\lambda x + \lambda E_{\mathbb{P}}\left(f^*\left(\frac{1}{\lambda}dQ/d\mathbb{P}\right)\right)\right)$, for $Q \ll \mathbb{P}$, and where $f : \mathbb{R} \to \mathbb{R}$ is convex and increasing and x is in the interior of $f(\mathbb{R})$, and f^* denotes its Legendre-Fenchel transform. The associated risk measure is given by $\rho(Y) = \inf\{m \in \mathbb{R} | E_{\mathbb{P}}(f(-Y-m)) \le x\}$, $Y \in L^\infty$, and is the well known shortfall risk measure introduced by Föllmer and Schied. Its dynamic version is weakly dynamically consistent but fails to be dynamically consistent.
- *Penalties associated with statistical distance functions:* $\gamma(Q) = E_{\mathbb{P}}(g(dQ/d\mathbb{P}))$, for $Q \ll \mathbb{P}$ and suitable functions g.

3.2. Forward robust portfolio selection. We consider the model as in [69] with $d+1$ securities whose prices, $(S^0; S) = (S_t^0, S_t^1, ..., S_t^d)$, $t \ge 0$, with $S_0 = 1$ (the numeraire) and S_t, $t \ge 0$, is a d-dim càdlàg locally bounded semimartingale on a complete filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,\infty)}, \mathbb{P})$. The wealth process is given by $X_t^\alpha = x + \int_0^t \alpha_s \cdot dS_s$, $t \ge 0$. The set $\mathcal{A}$ of admissible policies consists of weight portfolios α_t that are predictable and, for each $T > 0$ and $t \in [0,T]$, are S-integrable and $X_t^\alpha > -c$, $c > 0$. We denote the set of probability measures that are equivalent to $\mathbb{P}$ by $\mathcal{Q}$. For further details and all technical assumptions, see [69] and [26].

Definition 3.4.

i) A random field is a mapping $U : \Omega \times \mathbb{R} \times [0,\infty) \to \mathbb{R}$ which is measurable with respect to the product of the optional σ-algebras on $\Omega \times [0,\infty)$ and $\mathcal{B}(\mathbb{R})$.

ii) A utility field is a random field such that, for $t \ge 0$ and $\omega \in \Omega$, the mapping $x \to U(\omega, x, t)$ is $\mathbb{P}$-a.s. a strictly concave and strictly increasing $C^1(\mathbb{R})$ function, and satisfies the Inada conditions $\lim_{x\to-\infty}\frac{\partial}{\partial x}U(\omega,x,t) = \infty$ and $\lim_{x\to\infty}\frac{\partial}{\partial x}U(\omega,x,t) = 0$. Moreover, for each $x \in \mathbb{R}$ and $\omega \in \Omega$, the mapping $t \to U(\omega, x, t)$ is càdlàg on $[0,\infty)$, and for each $x \in \mathbb{R}$ and $T \in [0,\infty)$, $U(\cdot, x, T) \in L^1(\mathbb{P})$.

For simplicity, the $\omega-$notation is suppressed in $U(x,t)$. Next, the concept of an admissible penalty function is introduced.

Definition 3.5.

i) Let $T > 0$ and $t \in [0,T]$, and $\mathcal{Q}_T = \{Q \in \mathcal{Q} : Q|_{\mathcal{F}_T} \sim \mathbb{P}|_{\mathcal{F}_T}\}$. Then, a mapping $\gamma_{t,T} : \Omega \times \mathcal{Q}_T \to \mathbb{R}_+ \cup \{\infty\}$, is a penalty function if $\gamma_{t,T}$ is $\mathcal{F}_t$-adapted, $Q \to \gamma_{t,T}(Q)$ is convex a.s., for $Q \in \mathcal{Q}_T$, and for $\kappa \in L_+^\infty(\mathcal{F}_t)$, $Q \to E_Q(\kappa\gamma_{t,T}(Q))$ is weakly lower semi-continuous on $\mathcal{Q}_T$.

ii) For a given utility random field $U(x,t)$, $\gamma_{t,T}$ is an admissible penalty function if, for each $T > 0$ and $x \in \mathbb{R}$, $E_Q(U(x,T)) < \infty$ for all $Q \in \mathcal{Q}_{t,T}$, with $\mathcal{Q}_{t,T} = \{Q \in \mathcal{Q}_T : \gamma_{t,T}(Q) < \infty, \text{a.s.}\}$.

Using the above notions, the following definition of the *robust forward performance process* was proposed in [26]. Because of the presence of the penalty term in (3.7) below, it is more convenient to formulate this concept in terms of the self-generation property (cf. (2.13)).

Definition 3.6. Let, for $t \ge 0$, $U(x,t)$ be a utility field and, for $T > 0$ and $t \in [0,T]$, $\gamma_{t,T}$ be an admissible family of penalty functions. Define the associated value field as a family of

mappings $u(\cdot;t,T):L^\infty\to L^0(\mathcal{F}_t;\mathbb{R}\cup\{\infty\})$, given by

$$u(\xi;t,T)=ess\sup_{\pi\in\mathcal{A}_{bd}}ess\inf_{Q\in\mathcal{Q}_{t,T}}\left(E_Q\left(U(\xi+\int_t^T\alpha_s\cdot dS_s,T)\middle|\mathcal{F}_t\right)+\gamma_{t,T}(Q)\right), \tag{3.7}$$

with $\xi\in L^\infty(\mathcal{F}_t)$ and $\mathcal{A}_{bd}$ being the set of admissible policies in $\mathcal{A}$ that yield bounded wealth processes. Then, the pair $(U,\gamma_{t,T})$ is a forward robust criterion if, for $T>0$ and $t\in[0,T]$, $U(\xi,t)$ is self-generating, that is $U(\xi,t)=u(\xi;t,T)$, a.s..

Preliminary results for the dual characterization of forward robust preferences were recently derived in [26]. The dual of the utility field $U(x,t)$ is defined, for $(y,t)\in\mathbb{R}_+\times[0,\infty)$, as $\tilde{U}(y,t)=\sup_{x\in\mathbb{R}}(U(x,t)-xy)$. One, then, defines the dual value field, for $T>0$ and $t\in[0,T]$, as the mapping $\tilde{u}(\cdot,t,T):L^0_+(\mathcal{F}_t)\longrightarrow L^0(\mathcal{F}_t,\mathbb{R}\cup\{\infty\})$ given by

$$\tilde{u}(\eta;t,T)=ess\inf_{Q\in\mathcal{Q}_{t,T}}ess\inf_{Z\in\mathcal{Z}_T^a}\left(E_Q\left(\tilde{U}\left(\eta Z_{t,T}/{Z_{t,T}}^Q,T\right)\middle|\mathcal{F}_t\right)+\gamma_{t,T}(Q)\right). \tag{3.8}$$

Herein, $Z_{t,T}=Z_T/Z_t$ (resp. $Z^Q_{t,T}=Z^Q_T/Z^Q_t$), where Z_s (resp. Z^Q_s), $s=t,T$, is the well known density process for the absolutely continuous local martingale measures (resp. Q) (for further details, see [69]).

In turn, the pair $\left(\tilde{U},\gamma_{t,T}\right)$, for an admissible family of penalty functions $\gamma_{t,T}$, is said to be self-generating if $\tilde{U}(\eta,t)=\tilde{u}(\eta;t,T)$, for all $\eta\in L^0_+(\mathcal{F}_t)$. Under additional assumptions, it was shown in [26] that the primal and the dual value fields satisfy, for all $T>0$ and $t\in[0,T]$, the bidual relationships $u(\xi;t,T)=essinf_{\eta\in L^0_+(\mathcal{F}_t)}(\tilde{u}(\eta;t,T)+\xi\eta)$ and $\tilde{u}(\eta;t,T)=ess\sup_{\xi\in L^\infty(\mathcal{F}_t)}(u(\xi;t,T)-\xi\eta)$, for $\xi\in L^\infty(\mathcal{F}_t)$ and $\eta\in L^0_+(\mathcal{F}_t)$. It was also shown that the primal criterion $(U,\gamma_{t,T})$ is self-generating, and thus a forward robust criterion, if and only if its dual counterpart $\left(\tilde{U},\gamma_{t,T}\right)$ is self-generating.

There are several open questions for the characterization and construction of the robust forward performance process. For example, there are certain assumptions on $\mathcal{Q}_{t,T}$ in Definition 3.5 (see Assumption 4.5 in [26]) which might be difficut to remove. Another issue is whether the penalty functions need to be themselves dynamically consistent, in that whether they need to satisfy $\gamma_{t,T}(Q)=\gamma_{t,s}(Q)+E_Q(\gamma_{s,T}(Q)|\mathcal{F}_t)$, for $T>0$ and $t\in[0,T]$. As Definition 3.5 stands, this property is not needed as long as the pair $(U(x,t),\gamma_{t,T})$ is self-generating. However, examples (either for the primal or the dual forward robust criterion) for non dynamically consistent penalty functions have not been constructed to date. We remind the reader that classical robust utilities are well defined even if the associated penalties are not time-consistent, with notable example being the penalty associated with the shortfall risk measure. It is not clear, however, if in the forward setting such cases are indeed viable.

Because of the model ambiguity and the semimartingale nature of the asset prices, it is not immediate how to obtain the robust analogue of the forward performance SPDE (2.14). Some cases have been analyzed in [26]. Among others, it is shown that when asset prices follow Itô processes and the forward robust criterion is time-monotone, then its dual $\tilde{U}(x,t)$ solves a fully non-linear ill-posed PDE with random coefficients.

The time-monotone case with logarithmic initial datum, $U(x,0)=\ln x$, and time-consistent quadratic penalties can be explicitly solved. The optimal policy turns out to be a *fractional Kelly strategy,* which is widely used in investment practice. The fund manager invests in the growth optimal (Kelly) portfolio corresponding to her best estimate of the market

price of risk. However, she is not fully invested but, instead, allocates in stock a fraction α_t^* of her optimal wealth that depends on her "trust" in this estimate. Her "trust" is modelled by a process δ_t that appears in the quadratic penalty. As $\delta_t \uparrow \infty$ (infinite trust in the estimation), α_t^* converges to the classical Kelly strategy associated with the most likely model while if $\delta_t \downarrow 0$ (no trust in the estimation), α_t^* converges to zero and deleveraging becomes optimal.

4. Concluding remarks

Despite the numerous advances on the theoretical development and analysis of portfolio management models and their associated stochastic optimization problems, there is relatively little intersection between investment practice and academic research. As mentioned in the introduction, the two main reasons for this are the fundamental difficulties in estimating the parameters for the price processes and the lack of practically relevant investment performance criteria.

While estimating the volatility of stock prices is a problem extensively analyzed (see, for example, [2] and [47]), estimating their drift is notoriously difficult (see, among others, [17] and [39]). Note that drift estimation is not an issue in derivative valuation, for pricing and hedging do not require knowledge of the historical measure but, rather, of the martingale one(s). As a result, there is no need to estimate the drift of the underlying assets. Recently, a line of research initiated by S. Ross ([54]) on the so called *Recovery Theorem* investigates if the historical measure can be recovered from its martingale counterpart(s) (see also [10]).

The lack of a realistic investment performance criterion poses equally challenging questions. There are two issues here: the form of the criterion per se, and its dynamic and time-consistent nature. A standard criticism from practitioners is that utility functions are elusive and inapplicable concepts. Such observations date back to 1968 in the old note of F. Black ([5]). Indeed, in portfolio practice, managers and investors have investment targets (expected return, volatility limits, etc.) and companies have constraints on their reserves and risk limits, and it is quite difficult, if possible at all, to map these inputs to a classical utility function. The only criterion that bridges part of this gap is the celebrated *mean-variance* one, developed by H. Markowitz ([38]), which corresponds to a quadratic utility with coefficients reflecting the desired variance and associated optimal mean. However, this widely used criterion is essentially a single-period one. In a multi-period setting, it becomes time-inconsistent, in contrast to criteria used in derivative pricing which are by nature dynamically consistent. It is not known to date how to construct genuinely dynamic and time-consistent mean-variance or other practically relevant investment criteria. Some attempts towards this direction can be found in the recent works [48] and [68].

Acknowledgements. The author would like to thank B. Angoshtari and S. Kallblad for their comments and suggestions.

References

[1] Artzner, P., Delbaen, F., Eber, J.-M., and D. Heath, *Coherent measures of risk*, Mathematical Finance **9**(3) (1999), 203–228.

[2] Barndorff-Nielsen, O.E., and N. Shephard, *Econometric analysis of realized volatility and its use in estimating stochastic volatility models*, Journal of the Royal Statistical Society, Series B (2002), 253–280.

[3] Bayraktar, E. and M. Sirbu, *Stochastic Perron's method for Hamilton-Jacobi-Bellman equations*, SIAM Journal on Control and Optimization **51**(6) (2013), 4274–4294.

[4] Berrier F., Rogers L.C. G., and M. Tehranchi, *A characterization of forward utility functions*, preprint, (2009).

[5] Black, F., *Investment and consumption through time*, Financial Note No. **6B**, Arthur D. Little, Inc. (1968).

[6] Borkar, V.S., *Optimal control of diffusion processes,* Pitman Research Notes **203** (1983).

[7] Borrell, C., *Monotonicity properties of optimal investment strategies for log-normal Brownian asset prices*, Mathematical Finance **17**(1) (2007), 143–153.

[8] Bouchard, B. and H. Pham, *Wealth-path dependent utility maximization in incomplete markets*, Finance and Stochastics **8** (2004), 579–603.

[9] Bouchard, B. and N. Touzi, *Weak Dynamic Programming Principle for viscosity solutions*, SIAM Journal on Control and Optimization **49**(3) (2011), 948–962.

[10] Carr, P and J. Yu, *Risk, return, and Ross recovery*, Journal of Derivatives **20**(1) (2012).

[11] Crandall, M., Ishii, H., and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bulletin of the American Mathematical Society **27** (1992), 1–67.

[12] Detemple, J., Garcia, R., and M. Rindisbacher, *A Monte Carlo method for optimal portfolios*, Journal of Finance **58**(1) (2003), 401–446.

[13] Duffie, D. and P.-L. Lions, *PDE solutions of stochastic differential utility*, Journal of Mathematical Economics **21** (1992), 577–606.

[14] El Karoui, N., Nguyen, D.H., and M. Jeanblanc, *Compactification methods in the control of degenerate diffusions: existence of an optimal control*, Stochastics **20** (1987), 169–220.

[15] El Karoui, N. and M. M'rad, *An exact connection between two solvable SDEs and a nonlinear utility stochastic PDE*, preprint, arXiv:1004.5191, (2010).

[16] _____, *Stochastic utilities with a given portfolio: approach by stochastic flows*, preprint, arXiv:1004.5192, (2010).

[17] Fama, E.F. and K.R. French, *The cross-section of expected stock returns*, Journal of Finance **88**(5) (1980), 829–853.

[18] Fleming, W.H. and H.M. Soner, *Controlled Markov processes and viscosity solutions*, Springer-Verlag, 2nd edition (2006).

[19] Föllmer F. and A. Schied, *Convex measures of risk and trading constraints*, Finance and Stochastics **6** (2002), 429–447.

[20] Föllmer, H., Schied, A., and S. Weber, *Robust preferences and robust portfolio choice*, Handbook of Numerical Analysis **15** (2009), 29–87.

[21] Frittelli, M. and Rosazza Gianin, E., *Putting order in risk measures*, Journal of Banking and Finance **26** (2002), 1473–1486.

[22] Gilboa, I. and D. Schmeidler, *Maxmin expected utility with non-unique prior*, Journal of Mathematical Economics **18**(1989), 141–153.

[23] Heath, D., *Back to the future, Plenary Lecture*, First World Congress of the Bachelier Finance Society, Paris, 2000.

[24] Itô, K., *On stochastic processes. I.* (Infinite divisible laws of probability), Japanese Journal of Mathematics **18** (1942), 261–301.

[25] Kallblad, S. and T. Zariphopoulou, *Qualitative properties of optimal investment strategies in log-normal markets*, submitted for publication (2014).

[26] Kallblad, S., Obloj, J., and T. Zariphopoulou, *Model uncertainty, robust forward criteria and fractional Kelly strategies*, preprint (2013).

[27] Karatzas, I., Lehoczky, J.P., Shreve S., and G.-L. Xu, *Martingale and duality methods for utility maximization in an incomplete market*, SIAM Journal on Control and Optimization, **25** (1987), 1157–1586.

[28] Karatzas, I. and G. Zitkovic, *Optimal consumption from investment and random endowment in incomplete semimartingale markets*, Annals of Applied Probability **31**(4) (2003), 1821–1858.

[29] Kim, T.S. and E. Omberg, *Dynamic nonmyopic portfolio behavior*, Review of Financial Studies **9**(1) (1996), 141–161.

[30] Kramkov, D. and W. Schachermayer, *The asymptotic elasticity of utility functions and optimal investment in incomplete markets*, Annals of Applied Probability **9**(3) (1999), 904–950.

[31] ———, *Necessary and sufficient conditions in the problem of optimal investment in incomplete markets*, Annals of Applied Probability **13**(4) (2003), 1504–1516.

[32] Kramkov, D. and M. Sirbu, *On the two times differentiability of the value functions in the problem of optimal investment in incomplete market*, Annals of Applied Probability **16**(3) (2006), 1352–1384.

[33] Krylov, N., *Controlled diffusion processes*, Springer-Verlag (1987).

[34] Larsen, K. and G. Zitkovic, *Stability of utility-maximization in incomplete markets*, Stochastic Processes and their Applications **117**(11) (2007), 1642–1662.

[35] Lions, P.-L., *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations*, Part I: The Dynamic Programming Principle and applications, Communications in Partial Differential Equations **8** (1983), 1101–1174.

[36] _____, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations*, Part II: Viscosity solutions and uniqueness, Communications in Partial Differential Equations **8** (1983), 1229–1276.

[37] Maccheroni, F., Rustichini, and Marinacci, M., *Ambiguity aversion, robustness and the variational representation of preferences*, Econometrica **74** (2006), 1447–1498.

[38] Markowitz, H., *Portfolio selection*, Journal of Finance **7** (1952), 77–91.

[39] Mehra R. and E. Prescott, *The equity premium: a puzzle*, Journal of Monetary Economics **15**(2) (1985), 145–161.

[40] Merton, R., *Lifetime portfolio selection under uncertainty: the continuous-time case*, The Review of Economics and Statistics **51** (1969), 247–257.

[41] Monin, P., *On a dynamic adaptation of the Distribution Builder approach to investment decisions*, Quantitative Finance **14**(5) (2014).

[42] Monin, P. and T. Zariphopoulou, *On the optimal wealth process in a log-normal market: Applications to Risk Management*, Journal of Financial Engineering, in print (2014).

[43] Musiela, M. and T. Zariphopoulou, *Optimal asset allocation under forward exponential criteria*, Markov Processes and Related Topics: A Festschrift for Thomas. G. Kurtz, IMS Collections **4** (2008), 285–300.

[44] _____, *Portfolio choice under dynamic investment performance criteria*, Quantitative Finance **9** (2009), 161–170.

[45] _____, *Portfolio choice under space-time monotone performance criteria*, SIAM Journal on Financial Mathematics **1** (2010), 326–365.

[46] _____, *Stochastic partial differential equations in portfolio choice*, Contemporary Quantitative Finance (C. Chiarella and A. Novikov, eds.) (2010), 195–215.

[47] Mykland, P. and L. Zhang, *Inference for continuous semimartingales observed at high frequency*, Econometrica **77**(5) (2009), 1403–1445.

[48] Musiela, M., Vitoria, P., and T. Zariphopoulou, *Infinitesimal mean-variance, time consistency and convergence*, preprint (2014).

[49] Nadtochiy, S. and M. Tehranchi, *Optimal investment for all time horizons and Martin boundary of space-time diffusions*, submitted for publication (2013).

[50] Nadtochiy, S. and T. Zariphopoulou, *An approximation scheme for solution to the optimal investment problem in incomplete markets*, SIAM Journal on Financial Mathematics **4**(1) (2013), 494–538.

[51] _____, *A class of homothetic forward investment performance processes with non-zero volatility*, Inspired by Finance, 475–504, Springer (2014).

[52] Pham, H., *Smooth solutions to optimal investment models with stochastic volatilities and portfolio constraints*, Applied Mathematics and Optimization **46** (2002), 1–55.

[53] Quenez, M.-C., *Optimal portfolio in a multiple prior model. Random Fields and Applications IV*, Progress in Probability **58**, Birkhäuser, 291–321 (2004).

[54] Ross, S., *The Recovery Theorem*, Journal of Finance, (2013).

[55] Schachermayer, W., *Optimal investment in incomplete markets when wealth may become negative*, Annals of Applied Probability **11**(3) (2001), 694–734.

[56] _____, *A super-martingale property of the optimal portfolio process*, Finance and Stochastics **7**(4) (2003), 433–456.

[57] Schied, A., *Optimal investments for risk - and ambiguity - averse preferences: a duality approach*, Finance and Stochastics **11**(1) (2007), 107–129.

[58] Schmeidler, D., *Subjective probability and expected utility without additivity*, Econometrica **57** (1989), 571–587.

[59] Schroder, M. and C. Skiadas, *Optimal lifetime consumption-portfolio strategies under trading constraints and generalized recursive preferences*, Stochastic Processes and their Applications **108** (2003), 155–202.

[60] Talay, D. and Z. Zheng, *Worst case model risk management*, Finance and Stochastics **6** (2002), 517–537.

[61] Touzi, N., *Stochastic control problems, viscosity solutions and application to finance*, Lecture Notes, Scuola Normale Superiore, Pisa (2002).

[62] Wachter, J., *Risk aversion and allocation to long term bonds*, Journal of Economic Theory **112** (2003), 325–333.

[63] Widder, D.V., *The heat equation*, Academic Press (1975).

[64] Xia, J., *Risk aversion and portfolio selection in a continuous-time model*, SIAM Journal on Control and Optimization **49**(5) (2011), 1916–1937.

[65] Yong, J. and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer, New York (1999).

[66] Zariphopoulou, T., *A solution approach to valuation with unhedgeable risks*, Finance and Stochastics **5** (2001), 61–82.

[67] _____, *Optimal asset allocation in a stochastic factor model - an overview and open problems*, Advanced Mathematical Modeling, Radon Series in Computational and Applied Mathematics **8** (2009), 427–453.

[68] Zhou. X.Y. and T. Zariphopoulou, *Time-consistent dynamic Markowitz strategies*, work in progress (2014).

[69] Zitkovic, G., *A dual characterization of self-generation and log-affine forward performances*, Annals of Applied Probability **19**(6) (2010), 2176–2210.

[70] _____, *Utility theory - historical perspectives*, Encyclopedia of Quantitative Finance, R. Cont Ed., (2010)

[71] _____, *Dynamic Programming for controlled Markov families: abstractly and over martingale measures*, SIAM Journal on Control and Optimization, in print.

Depts. of Mathematics and IROM, The University of Texas at Austin, Austin, USA, 78712
E-mail: zariphop@math.utexas.edu

18. Mathematics Education and Popularization of Mathematics

The internet and the popularization of mathematics

Étienne Ghys

Abstract. In this paper, "popularization of mathematics" is understood as the attempt to share some of the current mathematical research activity with the general public. I would like to focus on the internet as a powerful tool to achieve this goal. I report on three personal experiences: the making of two animation films available on the web, the participation to a web-journal aimed at a wide audience, and the filming of a 15 minute video clip.

Mathematics Subject Classification (2010). Primary 00A09, 97A80; Secondary 97A40.

Keywords. Popularization of mathematics, internet.

1. Introduction

Even though the *International Congress of Mathematicians* has been devoting one of its sections to mathematical education for quite some time, the inclusion of "popularization" in its realm is rather recent. Only five talks discussed this topic in previous congresses [11, 20–22, 25]. Among these contributions, I would like to mention Ian Stewart's article which analyzes in depth the many possible types of media which can be used for popularization. He focuses on magazines, newspapers, books, radio and television but barely mentions the internet. Eight years later, the internet is unavoidable. It has changed our everyday life, be it private or professional. I am convinced that in 2014, the internet should be *the* main tool for the popularization of mathematics and that the mathematical community has the duty of learning how to use this incredible communication instrument. This is not easy and much remains to be done.

I would like to report on three very specific experiences in which I have been involved in recent years: the production of two mathematical films freely available on the web, the creation of a web-based journal aimed at a wide audience and the recording of a very short clip for the web. My intention is to illustrate some of the difficulties that mathematicians can encounter in these kinds of ventures and to propose possible improvements.

This paper is not an attempt to describe in a systematic way all the issues related to mathematics and the internet. My only purpose is to give an account of a very personal experience.

It is a pleasure to thank Jos Leys and Aurélien Alvarez for their collaboration, as well as all the members of the editorial board of *Images des Mathématiques*. I also thank Marie Lhuissier for her very helpful comments.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

2. Why popularization?

Amazingly, most articles related to popularization begin with a section trying to explain why this is a honorable occupation. A similar section in a paper dedicated to geometry or topology, for example, would seem inappropriate in the proceedings of the ICM. It is a fact that most of our colleagues are not convinced that popularization is a respectable mathematical activity. There is a need for justification.

My first comment would be that preparing any kind of "popular" presentation is a real challenge, and very frequently forces you to understand much better the topic you want to present: a profitable investment for mathematicians! In [13] Sir Christopher Zeeman explained that after delivering his *Christmas lectures* in 1978 [24], he received a message from the chairman of the British *Science Research Council* who "tickled him off for wasting his time popularizing on TV instead of doing research". Zeeman answered that these lectures had in fact inspired a research paper in dynamical systems.

Let me quote David Hilbert in the introduction of his famous lecture in Paris during the ICM 1900[12][1] .

> "A mathematical theory is not to be considered complete until you made it so clear that you can explain it to the man you meet on the street. For what is clear and easily comprehended attracts and the complicated repels us."

Moreover, again from an egocentric point of view, popularization (like teaching) is highly rewarding for the working mathematician. A typical research paper has a few dozens readers (in favorable cases) and this can be frustrating, but a good popular paper can easily be read by thousands of web-users.

At the wider level of mathematics as a whole, Lásló Lovász explains clearly the importance of communication [19]:

> "A larger structure is never just a scaled-up version of the smaller. In larger and more complex animals an increasingly large fraction of the body is devoted to "overhead": the transportation of material and the coordination of the function of various parts. In larger and more complex societies an increasingly large fraction of the resources is devoted to non-productive activities like transportation information processing, education or recreation. We have to realize and accept that a larger and larger part of our mathematical activity will be devoted to communication."

Note that this comment primarily applies inside mathematics, with all its subcultures in danger of blowing up into many disconnected components. But it also applies to the communication from inside to outside mathematics, which is the subject of popularization. The ever expanding mathematical body requires more elaborate and stronger links to remain connected to the society at large.

In the same paper, Lovász asks for a special training of our students:

> "While full recognition of expository work is still lacking, the importance of it is more and more accepted. On the other hand, mathematics education does little to prepare students for this. Mathematics is a notoriously difficult subject

[1] As a matter of fact, Hilbert quotes "un mathématicien français du temps passé" who seems to be Gergonne.

> to talk about to outsiders (including even scientists). I feel that much more effort is needed to teach students at all levels how to give presentations, or write about mathematics they learned. (One difficulty may be that we know little about the criteria for a good mathematical survey)."

From another point of view, the necessity of popularizing mathematics is a direct consequence of the significant decrease of the number of math students, or more generally of scientific students: it is therefore a matter of survival for our discipline. It is our duty to explain to the young generation why mathematics is so beautiful and gratifying, and can be a wonderful option for their careers. A few decades ago, the prestige of science in society was much higher and there was some kind of natural flux coming into mathematics.

Of course, one should emphasize that popularizing mathematics does not only consist in advertising academic careers and in producing more research mathematicians! There is also an obvious utilitarian economic issue since our contemporary world needs more scientists and therefore more mathematicians. If we want more engineers, scientists and mathematicians, we need a general population which is at least aware of the existence of mathematicians. A significant part of the population is indeed convinced that there is nothing more to do in mathematics, and that the field has been closed since the ancient Greeks. Somehow, the most important goal of popularization is not necessarily to convey a specific mathematical content, but to explain that math/science could be a real option for themselves, or for their kids, or at least to show that it is a respectable activity, useful for society at large. More than two thousand years ago, Archimedes wrote *Sand-Reckoner* as a letter to his powerful king. That was a way of expressing the necessity of science for his kingdom. Today, we do not care about kings, but taxpayers want to understand what we are doing with their money and they deserve candid answers.

One should of course not forget the cultural aspect of mathematics, so obvious for professional mathematicians and so unknown to the general population. We have to explain that it is important for the "man on the street" to have some taste for mathematics (or science in general) in the same way as, for instance, it is important to enjoy the arts. Such a taste is not necessarily related to the "usefulness" of mathematics, say for economics or engineering sciences, and does not require a deep understanding of technical details. One should make clear that mathematics can be fun and interesting to everybody, just as literature can be enjoyed at many levels.

The choice of popularizing science is clearly a political and democratic issue. As a historical example, in 1841 François Arago, then director of the Paris observatory, built a large lecture hall in the heart of the main building, entirely dedicated to his weekly lectures on "Popular Astronomy". These lectures, aimed at the general public, were indeed a great popular success (see the marvelous notes [2]). His successor, Urbain Leverrier, decided to transform the observatory into an efficient modern laboratory, fully devoted to research. He demolished Arago's lecture hall[2] :

> "The amphitheater is and will remain purposeless. The Observatory should not compete with the organizations of public instruction located in the very center of Paris, which suffice for their task. An institution which is requested to work at the progress of science [...] must look for the most absolute tranquility" [17].

Two great scientists and two different approaches to the relationship between science and

[2] and turned it into a private apartment for his personal use!

society.

For more on this topic, including a discussion on the history of popularization, I refer to [13].

3. The specificity of the internet

Of course, mathematics is already present at many levels on the internet. One finds thousands of blogs, some of them very popular among... professional mathematicians (for example Tao's blog) but most are not related to popularization. One also finds many websites of teachers sharing their enthusiasm for mathematics but they are usually connected to education and not to mathematical research. The Khan Academy provides a fantastic example of an internet access to education: it contains thousands a short clips covering mathematics from elementary to high school (and even some calculus). Wikipedia is an incredible success in general, and in mathematics in particular, but one should probably not qualify it as popularization. I would like to restrict myself here to websites dedicated to the presentation of some current mathematical developments to the general population (and therefore not aimed at professional mathematicians). Even with this restriction, one finds hundreds of websites, from individual blogs (for instance www.science4all.org) to institutional ones (among many more examples accromath.uqam.ca, plus.maths.org, maddmaths.simai.eu, interstices.info). Many institutions have subsections of their home pages devoted to outreach (for example, www.simonsfoundation.org).

The internet is an incredible jungle. Unlike mathematical papers or books, which are more or less built on similar structures, there is no unity on the web. The first mistake would be to try to export our professional habits and to produce webpages which look like mathematical books, with theorems and lemmas. A new tool should not be used to do what we have been doing for many years, even if we can do it faster or more easily : it should instead be used to do something new and more efficient.

Pictures, movies, music, podcasts or apps provide innovative and fascinating instruments to communicate mathematics, in a way which is very different from traditional texts. It is not the purpose of this paper to discuss the potential use of these new tools in research but I mention for instance that some online mathematical journals include short videos by the authors, presenting their own papers[3] .

In the domain of popularization, the possibilities are infinite and are still to be explored. As an example, one could easily break the traditional ordering in a mathematical text and let the reader-viewer-listener[4] choose his/her own trajectory inside a rich network of possibilities, according to his/her own background or taste, making him/her more of an actor than a passive reader. This may be the most important paradigmatic shift implied by the internet : from information organized in totally ordered lists to information located in a network. One could almost say that the information is not located on specific places but *coincides* with the network as a whole. A graph is much more than its vertices.

[3] Could a movie be considered as a bona fide proof of a theorem? Hilbert discusses the status of a picture: "The use of geometrical signs as a means of strict proof presupposes the exact knowledge and complete mastery of the axioms which underlie those figures; and in order that these geometrical figures may be incorporated in the general treasure of mathematical signs, there is necessary a rigorous axiomatic investigation of their conceptual content" [12]. For instance, the movie *Outside In* is very close to an actual proof of Smale's inversion theorem [18].

[4] The internet does not give access to smell, taste and touch... so far!

One should realize that when we surf the internet, we hop from webpage to webpage and usually spend a very short time on a given page. The typical "bounce rate" of a website is about 1/2: after viewing the entry page, half of the visitors immediately go somewhere else. Also, web-users do not read linearly, from top to bottom. One could argue that similar facts also apply to mathematical books or papers and that nowadays most of us "read" dozens of preprints at the same time, hopping from theorem to theorem, in the hope of finding something that could be useful for our research. However, the two hopping styles are very different. We should study and understand much better this new reading style on the web, closer to a random walk in a graph than to a motionless scholar reading in a library.

A related aspect of the internet, which is *a priori* in contradiction with the spirit of mathematical research, is its incredible speed and reactivity. Mathematicians usually spend months (or years) writing papers which will be read by a handful of people while web-users spend a few minutes posting tags with an improbable spelling on their *Facebook Wall*. Clearly these are two different communication modes and we should be able to switch from one to the other, keeping in mind their advantages and drawbacks. Inside the realm of mathematical research, nobody would deny the fundamental importance of long, difficult and carefully written papers. This requires time and is not compatible with "speed science". At some other moments, the researcher needs a quick answer to a specific question and he or she can frequently get immediate answers from MathOverflow : the "blog" style is efficient in these cases.

Is "speed science" compatible with popularization? Does it make sense for graduate students to participate in tournaments like the Three Minute Thesis competition? Even though most of us are reluctant to work at such a speed and look for peace, the answer to these questions has to be yes, if we do not want to lose contact with the younger generation. More importantly, in many cases (but nor all), I believe that a good popularization can be speedy, especially when the expected public has no connection at all with mathematics.

Another important aspect that makes the internet different is related to the validation problem. Everything can be posted on the internet, the best and the worst. No "referees" are present to prevent mistakes. Very often the general public would like to get some kind of certification that the content of a webpage is valid. This should be the role of mathematicians and we have to be creative in this respect. Can we trust the "wisdom of the crowds" and promote some verification in which everyone is encouraged to participate, in the spirit of *Wikipedia*? On the contrary, should we "export" some of our traditional refereeing methods based on anonymity?

The internet is the kingdom of wild plagiarism. It is amazing to see how a given text can travel from place to place, often subject to various "simplifications" or "additions", frequently with no mention of the original author. Mathematicians should understand that it is in some sense a great honor that their contributions are "duplicated" in many places. Of course, ideally, this should be done under the control of the author, but it is much better to accept it as a rule of the game. Trying to prevent this natural diffusion would be fighting a rearguard battle.

All these apparent drawbacks should be seen positively as powerful new opportunities. The ability to get information on almost any aspect of knowledge within a few clicks is of course a revolution. Older mathematicians remember their endless searches in libraries, going through the many (paper) volumes of *Mathematical Reviews*. Today, the published

mathematical literature is easily available[5] and arxiv.org provides access to preprints in real time. This high connectivity did not only change the everyday life of researchers. Amateurs surfing the web can now find quickly all kinds of information, for example on popular mathematics... if we know how to create easily accessible quality websites.

In a nutshell, the internet is working in a way which may not always look compatible with our tradition. We have to adapt and to learn how to play this new wonderful instrument.

4. First example: *Dimensions* and *Chaos*

4.1. Genesis of the project. In 2006, as I was preparing slides for a general public talk [6], I wanted to use some mathematical images that I liked on the website www.josleys.com. I therefore asked the webmaster for permission to use them. After my talk, I thanked him and asked for more information concerning his website. Jos Leys is a mechanical engineer who recently retired from a major chemical company. "At last, I can do mathematics!", he told me... Jos' mathematical background is typical for an engineer trained forty years ago: he had mastered pretty well classical analytic and differential geometry, but of course has no idea of contemporary mathematics at a research level. However, he has been interested in fractal geometry and computers since the early 80's. He genuinely loves mathematics. An article in *Pour la Science* portrayed him as an artist-geometer. At the same time, I was preparing a plenary lecture for ICM 2006 and my intention was to present, among other things, a result connecting periodic orbits in the Lorenz attractor and closed geodesics on the modular surface [7]. To my mind, this was a very visual theorem, but I did not know how to transform in practice my imprecise mental images into actual images. I therefore asked Jos for help in producing pictures. We did produce beautiful pictures, some of them being rather intricate, in particular those related to modular forms. Quickly, we realized that in order to explain ideas from dynamical systems, it was in fact best to use pictures in motion: movies! I was quite satisfied with the result and about one third of my talk in Madrid turned out to consist of movies. After the talk, Jos told me: "Now you have to explain to *me* the meaning of the movies *I* prepared with you". I was facing Hilbert's challenge: to make it so clear that you can explain it to the man you meet on the internet.

We first wrote some kind of "visual article", including movies, that we published in the web *Feature column* of the AMS [10]. However, this was not aimed at a "popular level" and Jos wanted something much more elementary. For instance, it was not possible to use complex numbers without explaining what they are... We therefore decided to produce a fully fledged film from scratch, starting at a very elementary level and, hopefully, going to our target: periodic orbits of the Lorenz attractor and closed geodesics on the modular surface. We were very optimistic but we quickly realized that it was not realistic in a single film. Soon, Aurélien Alvarez, who was at the time a graduate student, joined our team. So far, we "only" have produced parts 1 and 2, each two hours long, of a saga which could very well turn out to be infinitely long.

Part 1 is entitled "*Dimensions*". Its main purpose is to provide an introduction to dimension 4. More precisely, it gives a presentation of the 3-sphere inside 4-space and of the Hopf fibration.

Part 2 is entitled "*Chaos*". It is an elementary introduction to dynamical systems. The

[5] I don't comment here on the price of mathematical journals.

final chapters try to give a very rough idea of current conjectures on the statistical theory of strange attractors, like the Lorenz butterfly.

We are still far away from the modular surface and its geodesics!

4.2. The making of *Dimensions*. *The first decision* was to produce a film that would be split into "chapters", each being 13 minutes long (which is some kind of time unit in the video world). These chapters had to be as independent of each other as possible, and the mathematical level had to be increasing. Chapter 1 should be understandable by young children and the final chapters by undergraduates. The main idea was to propose to the spectator some kind of menu in which (s)he can select what (s)he wants. Some would only look at the first two chapters, others would only look at the last two and some would only look at chapters 5 and 6, for instance. Of course, this necessitated the careful writing of a scenario, in such a way that the many subsets look (and are) coherent. It would be frustrating for a spectator to see a film which leads him/her to a final chapter which is not understandable to him/her.

Here is the structure of the first movie *Dimensions*.

- *Chapter 1 (dimension two)* is very elementary. It contains the description of the 2-sphere in space, with its parallels and meridians, and shows the stereographic projection.
- *Chapter 2 (dimension three)* is still elementary and is based on the famous popular novel *Flatland* [1].
- *Chapters 3 and 4* get into the fourth dimension. They rely heavily on regular polytopes in dimension 4, seen as drawn on the 3-sphere, and then projected stereographically on 3-space (and then on the 2-dimensional computer screen).
- *Chapters 5 and 6 (complex numbers)* contain a visual introduction to complex numbers. These chapters are completely independent from the others and have been used quite a lot in classrooms.
- *Chapters 7 and 8 (Hopf fibration)* are the hardest parts. We show the linking of Hopf circles and the wonderful Villarceau circles on tori of revolution.
- *Chapter 9 (proof)* is special. It contains the complete proof that the stereographic projection maps circles to circles (or straight lines). This proof uses nothing above the level of secondary school, and we could very well have put this chapter right after chapter 1. We wanted to explain that mathematics is above all a matter of proofs, not only pictures.

For example, we propose the following combinations of chapters: Junior High School (1 or 1-2 or 1-2-9), High School (1-2-3-4-9, or 5-6), Undergraduates (2-3-4-5-6 or 5-6-(7-8-9)), College (7-8), General public (1-2-3-4).

The second decision was to tell a story. Each chapter is "presented" by a famous mathematician, from Hipparchus (for chapter 1), to Heinz Hopf (describing his fibration), along with Adrien Douady (explaining complex numbers). It is well known that the rich and long history of mathematics is a powerful vector for popularization. Naturally, the scenario is not written as a course, in any sense of the term. For instance, our presentation of complex numbers is not intended as a substitute to some kind of tutorial. Many teachers have used it in their classes as a complement or sometimes as an introduction. We explain the general idea

of complex numbers, we show their geometric meaning (which unfortunately disappeared from many high school curricula), we deform (conformally!) the portrait of Douady, and we finally illustrate these notions with beautiful pictures of the Mandelbrot set. We try to be precise but never formal. The commentaries and the images are of course supposed to be understandable but we are aware of the fact that some spectators get lost along the way. In this (unwanted but likely) case, the film should be attractive enough to keep the attention.

Technically, *Dimensions* is an animation movie. Most of the 185 000 images have been produced using the (free) software PovRay. This is of course a huge amount of work. *Dimensions* was released in 2008, after 18 months of elaboration.

We quickly realized that many fellow mathematicians were happy to help, in many ways. For instance, we could provide subtitles in 20 languages and soundtracks in 8 spoken languages. The concept of mathematical community is not an abstraction!

We also developed a website www.dimensions-math.org (also in many languages), giving extra information and references.

4.3. The economic model. We believe that *mathematical popularization should be freely accessible on the web*. We therefore decided that all movies could be freely downloaded on our website, under a Creative Commons licence. As a result, we were happy to see that the movies quickly could diffuse all over the web, primarily on *YouTube*.

We also produced a DVD that is sold on the website at a nominal price. This is a non profit activity and all benefits are immediately "invested" to offer DVDs to some organizations (like for instance the *International Mathematical Olympiads*, or *MathEnJeans*, etc.).

4.4. *Chaos*. Our second movie *Chaos* was released in January 2013 and is based on the very same model. We tell the story of dynamical systems, going slowly from periodic motions and limit cycles to chaotic examples, including Smale's horseshoe and the Lorenz attractor.

- *Chapter 1 (Motion and determinism)* is a non technical preview of the whole story, explaining determinism, sensitivity to initial conditions, and giving a hint that one could understand chaotic systems through statistical methods.
- *Chapters 2 and 3 (Vector fields, and Mechanics)* are very basic and can be used in the classroom: they give a very quick introduction to velocity, acceleration and forces. They are independent from the other chapters.
- *Chapter 4 (Oscillations)* gives an introduction to limit cycles.
- *Chapters 5, 6, 7 (Billiards, Horseshoe, Lorenz butterfly)* describe three historical examples of chaotic behavior.
- *Chapters 8, 9 (Statistics, Chaotic or not?)* introduce to the concept of physical measure (Sinai-Ruelle-Bowen) in a very intuitive way and to the general conjecture of Palis describing the statistical behavior of a typical dynamical system.

We could benefit from help not only from friends in the mathematical community all around the world, but also from a famous French actor and Brazilian singer[6] , who dubbed the commentaries!

[6] Thierry Lhermitte and Thalma de Freitas.

4.5. Assessment. Of course, I would not report on these movies if I were not convinced that this turned out to be a success. It is difficult to quantify the number of viewers or even of downloads. The website *Dimensions* has five mirrors (in Beijing, Mexico, New York, Rio and Tokyo) and the only objective data is that they had more than two million unique visitors, from *all* countries in the world. Obviously, none of my previous productions has been so widely distributed and it was a real pleasure for us to receive congratulations from kids in the middle of China.

We received thousands of emails thanking us for our work, and asking for more. It is not easy to get some clear view of our audience from these emails since their diversities is very impressive, from very young children to people seing improbable connections between the fourth dimension and spirituality... Nevertheless, one could say that many viewers are amateurs in a way or another. They probably found on the web the popular mathematics that they were looking for.

Did we only reach amateurs who were already convinced? We did not have clearly in mind this "target" when we started the project. Clearly, amateurs should not be neglected and one should carefully analyze their requirements. However, the public of those who have no connection at all with mathematics is probably more important and requires a specific approach, with a much weaker mathematical content.

As for the DVD's, we produced 20 000 copies which have been either sold or offered. I am convinced that our choice of *Creative Commons* was the right decision and that no other economic option would have generated such a diffusion for mathematical movies. According to a private publisher that we have contacted at the beginning of the project, there is no market for this kind of film.

From the non positive side, it is clear that a two hour film entirely produced by three persons, with no budget, cannot be compared with a *Pixar* production. Obviously, it is the work of amateurs, with many drawbacks, especially related to the rhythm, which is sometimes too slow. Another difficulty is that we should have planned the scenario and the storyboard in their smallest details before starting the production of the first chapters. It is unclear whether it would have been more efficient to develop a much more expensive project and to involve professionals: this would have implied too much of a burden and would have hidden what drives much of us: the fun of doing mathematics.

A successful aspect of the films is the splitting into individual chapters which are more or less independent and can be combined in many possible paths, depending on the viewer. This has been appreciated. However, we have to admit that we did not use the full flexibility of internet. It would have probably been more efficient to produce something more interactive, in which the web-user could make more choices, in the spirit of video games. Of course, this would have been technically much more difficult, probably beyond our capabilities.

One could probably assert that *Dimensions* and *Chaos* deal with mathematics which are easy to popularize: topology, geometry and dynamics. It would be clearly more difficult to produce a film on algebra, number theory or modern algebraic geometry. In these cases, one should choose other internet tools. Even so, it is possible that some domains cannot be shared with the general population. However, this may not be a serious problem. Many aspects of astronomy for instance are too technical to be presented to a wide audience, but astrophysicists have enough beautiful pictures or fascinating stories to popularize their discipline in an exceptional way.

5. Second example: *Images des Mathématiques*

5.1. Genesis of the project. In the 1980's, the French *Centre National de la Recherche Scientifique* (CNRS) decided to publish, once every two years, a volume entitled *Images des Mathématiques* (IdM for short). The idea was to include a dozen articles giving some illustration of recent mathematical progress. The target of this booklet was not clearly defined but instructions were given to the authors that they should not write for their colleagues. A small number of issues appeared but the publication stopped very quickly. This publication was expensive, the published articles were in practice only readable by colleagues, and the 7 000 copies were very badly distributed.

In 2004 and 2006, Jacques Istas and myself edited two more volumes... with the same weaknesses. We realized that many of the printed copies did not go out of the strict circle of mathematical researchers and even that many were not opened at all... Even worse, most articles were not understandable by mathematicians from outside the field of the author. This was a waste of money and energy.

We decided to create a web journal, still hosted by the CNRS, with the same title, dedicated to explaining current mathematical research outside of the circle of research mathematicians, if possible to Hilbert's "man on the street". The main idea was to ask for the help of many colleagues and to create a large editorial board. This would provide an analogue of a daily newspaper, giving "news from the mathematical community" as often as possible, ideally daily... Five years after the opening, in January 2009, about 2000 articles have been published (see below).

Of course, this initiative is not isolated. In 2008, IMU and ICMI commissioned a project to revisit the intent of Felix Klein when he wrote "Elementary Mathematics from an Advanced Standpoint" one hundred years earlier [14]. As explained by the Klein committee: "The aim is to produce a book for upper secondary teachers that communicates the breadth and vitality of the research discipline of mathematics and connects it to the senior secondary school curriculum. The 300-page book, prepared in more than 10 languages, will be written to inspire teachers to present to their students a more informed picture of the growing and interconnected field represented by the mathematical sciences in today's world. We expect this will be backed up by web, print, and DVD resources." See the website blog.kleinproject.

As one can see, the expected audience of IdM is slightly different since the Klein project is written for teachers. Moreover, the Klein project is more thought as a data base than as a magazine giving information at a continuous pace.

5.2. Structure of IdM. IdM is organized like any research mathematical journal. The editorial board consists of about twenty mathematicians, each being in charge of some section of the journal (see this page). In turn, each section has its own sub-committee taking all editorial decisions relative to this section. The union of the editorial board and all subcommittees contains about sixty colleagues. As examples of sections: history, conjectures, current research, press review etc.

IdM publishes two kinds of contributions, *articles* and *columns*.

Articles are close to research papers in the sense that they are evaluated in a process which is similar to the standard refereeing system. When an article is submitted for publication (authors are almost always invited to contribute by a member of the board), it is deposited on a private page. A few hundred volunteers have agreed to read and comment papers before publication. A dozen of these volunteers are selected for each submitted article and they

have access to the private page containing the draft of the paper. Typically, one half of these "referees" are professional mathematicians. These referees can comment the paper in a forum accessible to the author, to the other referees, and to the editors. Note in particular that the referees are not anonymous, even though some of them are only identified through a pseudonym. The process of evaluation then takes the form of a "conversation", through this forum, between the author and the referees, and this implies a continuous change of the text. When the editor in charge considers that the paper is ready, it can be published. Typically, this process takes about two months. About one thousand such articles have been published in the last five years.

Most articles are original and have been written for IdM. The few exceptions are related to some partnerships with some other journals, agreeing to share some papers. I mentioned earlier the "plagiarism" question. Many blogs do not hesitate to copy parts of articles published elsewhere. Of course, one should criticize this behavior if the original author is not mentioned. However, I am in favor of the idea that a given article might be published in different places, in different forms, for different publics, preferably with the agreement and participation of the author.

Columns are much shorter and usually with much lighter mathematical content. This is somehow the blog part of IdM. A certain number of colleagues have agreed to be columnists and they are encouraged to publish short contributions, of course related to mathematics, but typically from a different point of view. This could be for instance a political opinion, or the review of a book, of a movie, or even a joke... Of course, these columns are not refereed but a small team checks them before their (quick) publication. IdM has now published about one thousand of these columns.

The question of the nature of the public is of course fundamental. IdM is in principle aimed at the general public but clearly a significant part of our readers *are* mathematicians. Many are teachers or students, or have some relationship with mathematics, so that they are mathematicians in some way or another. One of the main difficulties is to ignore research mathematicians, since IdM is not for them! The idea would be to propose something widely accessible (to French readers) but it is of course impossible to write texts which are suitable for *everybody*. We adopted a code inspired by the ski slopes rating colors, from the easy green slope to the black one, and even off-piste. The green slope requires in principle no knowledge in mathematics.

From the financial point of view, IdM is almost cost-free and receives a modest support from CNRS.

5.3. Assessment of IdM. The audience of IdM (as measured with *Google Analytics*) has been steadily increasing since the opening of IdM (with a quasi-periodic modulation, related to weekends, vacations etc.). Today, IdM receives about 4000 visitors a day. This is much less than what we would expect but one should keep in mind that this web journal is only available to French speaking readers (although the project of translating into Spanish is on schedule).

The main difficulty encountered by IdM is to find authors. As a rule, authors are mathematicians and not journalists. Most of our colleagues are under a publication pressure for their own career and, unfortunately, this kind of article is not yet considered valuable enough to be included in their publication list. A possible improvement, giving value to these popularization articles, would be to include them in databases, like *MathSciNet* or *Zentralblatt*[7]

. Indeed, from my own experience, the refereeing process in IdM is far more advanced than in most "standard" research journals.

Moreover, potential authors quickly realize that writing such articles is far from easy and requires a lot of work. More often than not, they have great difficulties in understanding that most of the words that they use daily are simply not in the vocabulary of the potential readers. Most mathematicians have a totally wrong idea of the mathematical knowledge of the general population. It is clearly difficult to explain a recent mathematical idea to "the man you meet on the street" and even sometimes it may be impossible. The main comment from non-mathematicians about articles from IdM is: "too complex and too long". Our community has to train students in this kind of exercise and this should be included in university curricula. Somehow, one could think of IdM as some kind of laboratory where we practice and improve our ability to write such papers.

One could reasonably question the fact that the authors of IdM are not journalists. Of course, journalists usually know their readers much better than mathematicians do. However, they (usually) do not know mathematics as we know it, from inside. I am convinced that the popularization of mathematics should not be *entirely* delegated to journalists. It is the duty of mathematicians to spread mathematics in the general public. See the article by M. Emmer on the relationship journalists-mathematicians, in [13].

The "semi-public" refereeing system works rather well. As described above, it involves a dozen volunteers for each article who share with the author a private forum. Almost always, the published paper is significantly different from its original version. Professional mathematicians are used to the "dry style" of referees reports. Sometimes, comments from professionals on articles submitted to IdM are expressed so strongly that the non professionals are impressed and hesitate to give their own opinion and remain silent. Usually, non professionals would like to say "I don't understand" and professionals "You forgot to add such and such theorems". As for the authors it is not uncommon that they have difficulties accepting comments on their papers by "referees" who are not experts, even though they represent a good sample of their readers.

Of course visitors are welcome to add comments at the end of articles, after publication. However, we noticed some rather surprising behavior on the part of the readers. Many hesitate a lot before posting a comment by some kind of self censorship. They seem to be "impressed" by the expertise of some authors.

We conducted a survey to get a better understanding of our readers. As we could imagine, a significant minority of our visitors consists of researchers in mathematics. A majority are teachers or students. We still do not reach the very young. Clearly the articles are too long and too difficult. Sadly, it should be noted that 80% of our visitors are male.

Another difficulty is related to the navigation inside IdM. We should use all the possibilities of the internet in order to propose multiple choices to our readers. Unfortunately, most visitors do not understand that behind the home page, there is a large data base of articles. We need keywords, tags and all sorts of modern navigation tools. A web designer is currently analyzing the structure of the "back office" of IdM and will propose solutions. This has of course a cost.

Even though there is still a lot of progress to be made, collaborating with the editorial board of IdM is a challenging and exciting experience.

---

[7] As of today, the administrators of these two databases have not answered our proposal for reviewing articles from IdM.

6. Third example : popular lectures, *les Ernest*

The idea of popular science lectures is certainly not new. For instance, in 1825 Michael Faraday inaugurated the Royal Institution Christmas Lectures aimed at a "juvenile auditory". Since 1967, they are broadcast on the BBC television network and they are very successful. One had to wait until 1978 before one of these series could be dedicated to mathematics (by Christopher Zeeman [24] and Marcus du Sautoy in 2006 [4]).

Nowadays, it has become fashionable for many mathematics departments or institutions to organize popular lectures. It is even common to include them in the program of scientific meetings, including the ICM. The main problem, not always understood by the organizers, is to define the public as clearly as possible and to make sure that it comes! It is impossible for the speaker to prepare a lecture if he or she does not know whether the audience will be "juvenile" or "retired" or consisting of professional mathematicians. All these publics are interesting but very different... Suppose for example that the speaker plans to explain that $\sqrt{2}$ is irrational and discovers that all spectators have a PhD in mathematics. I have personally had several bad experiences of this kind that I will not describe here.

It has also become usual to film these lectures and to post them on the internet. In many cases, the result is a disaster. As explained earlier, the internet is not a new tool for doing what we have been doing for many years. A mathematical lecture filmed with one fixed camera, with no film editing, can be very useful for research mathematics but is certainly not adapted to a popular presentation of mathematics. One problem is the length. Frequently, a live lecture in front of an active public can last one hour and still be a great success. The same lecture posted on the internet will have a very different reception. The web-viewer can (and probably will) hop to some other place with one click. Looking at a static blackboard on a screen quickly becomes boring unless this is a technical research talk and you are really interested in a proof.

One of the standard mistakes from the organizers is to inform the speaker that his/her talk will be recorded one second before the start of the lecture. Theater and cinema are certainly different activities.

For the internet, it is fundamental to enable the spectator to see many different aspects of the lecture. There should be a subtle balance between views of the speaker, of his/her slides, and of the public in the room. This implies a serious editing of the film and a competent technical staff. Everything should be prepared well in advance, in coordination with the speaker.

I would like to report on two personal examples that were quite successful. I gave a public lecture in 2010 in Paris, on the occasion of the Clay Conference in honor of the proof of the Poincaré conjecture [8]. The conditions were optimal: the wonderful amphitheater of the Institut d'Océanographie, a public of high school students (and some distinguished colleagues on the first row), and above all the very professional editing by François Tisseyre, who has a long experience in filming mathematics (see for instance [3]). However, even though the editing seems to me very good, I do not think that the video is adapted to the internet: too long and not directly intended *for* the web.

Les Ernest is an association of young students from the École Normale Supérieure of Paris[8] . They understood that the internet is not just a way of broadcasting standard lectures.

> "One ambition : to offer a format for lectures adapted to the new media. [...]

[8] Les Ernest is a nickname for the goldfish swimming in a pond of the ENS.

> Knowledge should be shared democratically. More than ever, new approaches, frequently interdisciplinary, are necessary to understand our world. Usual lectures are not compatible with the internet code."

Les Ernest are producing films which are very short : 15 minutes. They cover all kinds of subjects, but they seem to have hesitated to include a lecture on mathematics, since I recorded the first one (after a computer scientist) in 2014 [9]. These clips are primarily intended for the internet. However, the organizers are convinced that it is important for the speaker to have a public in front of him or her, but only as a motivation. For instance, the lights are oriented in a way which enables special effects on the web, even though it implies that the speaker barely sees the spectators. The staff uses an impressive number of cameras and they work very hard on the editing. More importantly, they prepare the lecture in advance with the speaker, give him/her useful tips, and describe in great detail the targeted audience. A collaboration between the speaker and the organization team is maybe the key to success.

One of the difficulties with a 15 minute film is that it is short ! We have to know exactly what to say and, above all, what not to say. Should one prepare a detailed speech in advance? I fear that most mathematicians are not actors and this would lead to an artificial tone. We should certainly not improvise in such circumstances. I believe one should prepare some kind of rather precise framework, containing some key sentences, and, of course, rehearse several times in front of a clock.

This association is very close in spirit to the TED Conferences (Technology, Entertainment, Design) which also contain a relatively small number of mathematics lectures. As two model examples of short popular internet lectures, I would recommend [5] and [23]. Note in particular that in these examples, the speakers do not go into any mathematical detail, but both do give a fairly good image of the role of mathematicians.

All these are one-shot videos and one could wonder whether one should not prepare popular internet lectures as one produces a movie, filming many more rushes than necessary for the final product, and spending most of the time in the editing. Again this is the difference between theater and cinema.

7. Some conclusions and suggestions

Among the many possible communication tools that can be used for popularizing mathematics, the internet is probably the most powerful and efficient. A single individual or a very small group of mathematicians can produce webpages which can be viewed by many web-users, at almost no cost.

We have to learn the language which is adapted to this media and which is very different from the traditional language in mathematics: different in speed, depth and length. The point is not to transmit everything about mathematical research, but something about it. Sometimes, it is even sufficient to transmit *nothing* besides the fact that there exists a very active field of research called mathematics.

The most important mistake that should be avoided is to do on the internet what we are used to do in papers, books, classrooms, lecture halls etc. The internet enables us to develop new concepts.

We have to train the younger generations of mathematicians in these techniques. Almost every mathematician should have some training but we should also encourage some students

to specialize in popularization. More importantly, we should consider them as colleagues, with a well defined field of expertise, just like algebraists, geometers or analysts, and we should not consider them contemptuously as "mere journalists".

This implies that popularization has to be evaluated in a rigorous way, just as research papers are refereed. Two centuries ago, the mathematical community was able to develop a system of journals, some of them being specialized, whose "qualities" can be (more or less) compared. There is a need for the creation of mathematical journals specializing in popularization, following strict validation criteria for the acceptance of their published "papers". This will not be easy, since indeed, these papers are never printed on paper... and can take many different forms, far away from our usual introduction-theorem-lemma-proof-conclusion mathematical "literature".

These journals should be considered as "standard" mathematics journals, indexed by the main data bases, supported by the national mathematical societies etc. Published papers should appear proudly in the CVs of mathematicians and should be taken into account by the various hiring or promotion committees.

In short, a mathematician answering the traditional question from a colleague "What's your field?" should not feel anymore ashamed when he or she replies "I work on popularization of mathematics".

En passant, note that almost all references below are freely available on the web...

References

[1] Abbott, E., *Flatland, a romance of many dimensions*, 1884.

[2] Arago, F., *Astronomie populaire*, 1854.

[3] Douady, A., La dynamique du lapin, (directors D. Sorensen, F. Tisseyre), Video, Écoutez voir, 1996.

[4] du Sautoy, M., *Royal Institution Christmas Lectures 2006: The Num8er My5teries*.

[5] ———, *Symmetry, reality's riddle*, Ted lecture 2009.

[6] Ghys, E., *Poincaré et le monde non euclidien*, Bibliothèque Nationale de France, March 2006.

[7] ———, *Knots and dynamics*, Video, Proceedings of the ICM 2006, Madrid.

[8] ———, *Les maths ne sont qu'une histoire de groupes*, Colloque Clay, Paris, 2010.

[9] ———, *Et si le théorème de Pythagore n'était pas vrai*, Les Ernest, 2014.

[10] Ghys, E. and Leys, J., *Lorenz and modular flows: a visual introduction*, AMS Feature Column, 2006.

[11] Hansen, V.L., *Popularizing Mathematics: From Eight to Infinity*, Proceedings of the ICM 2002, Beijing.

[12] Hilbert, D., *Sur les problèmes futurs des mathématiques*, English version, Bulletin of the American Mathematical Society, vol. 8, no. 10 (1902), 437–479, Proceedings of the ICM 1900, Paris.

[13] Howson, A.G. and Kahane, J.P., *The popularization of mathematics*, ICMI Studies, 1990.

[14] Klein, F., *Elementary mathematics from an advanced point of view*, MacMillan and Co, 1932.

[15] Leys, J, Ghys, E., and Alvarez, A., *Dimensions, a walk through mathematics*, 2008.

[16] ———, *Chaos, a mathematical adventure*, 2013.

[17] Lequeux, J., *Le Verrier - Magnificent and Detestable Astronomer*, Springer 2013.

[18] Levy, S., Maxwell, D. , Munzner, T., and Thurston, W., Outside In. AK Peters, Wellesley, MA, 1994. Video (21 min), Part I and PartII.

[19] Lovász, L, *Trends in Mathematics: How they could Change Education?*

[20] Rousseau C., *The Role of Mathematicians in the Popularization of Mathematics*, Proceedings of the ICM 2010, Hyderabad.

[21] Schneider, J., *Issues for the popularization of mathematics*, Proceedings of the ICM 1994, Zürich.

[22] Stewart, I., *Mathematics, the media, and the public*, Proceedings of the ICM 2006, Madrid.

[23] Villani, C., Ted X Lecture, 2011.

[24] Zeeman, C., *Royal Institution Christmas Lectures 1978: Mathematics into Pictures*.

[25] Ziegler, G., *Communicating Mathematics to Society at Large*, Proceedings of the ICM 2010, Hyderabad.

UMPA, CNRS ENS Lyon, 46 Allée d'Italie 69340 Lyon, France
E-mail: etienne.ghys@ens-lyon.fr

Teaching and learning "What is Mathematics"

Günter M. Ziegler and Andreas Loos

Abstract. "What is Mathematics?" [with a question mark!] is the title of a famous book by Courant and Robbins, first published in 1941, which does not answer the question. The question is, however, essential: The public image of the subject (of the science, and of the profession) is not only relevant for the support and funding it can get, but it is also crucial for the talent it manages to attract — and thus ultimately determines what mathematics can achieve, as a science, as a part of human culture, but also as a substantial component of economy and technology.

In this lecture we thus

- discuss the image of mathematics (where "image" might be taken literally!),
- sketch a multi-facetted answer to the question "What is Mathematics?,"
- stress the importance of learning "What is Mathematics" in view of Klein's "double discontinuity" in mathematics teacher education,
- present the "Panorama project" as our response to this challenge,
- stress the importance of *telling stories* in addition to *teaching* mathematics, and finally
- suggest that the mathematics curricula at schools and at universities should correspondingly have space and time for at least three different subjects called Mathematics.

Mathematics Subject Classification (2010). Primary 97D30; Secondary 00A05, 01A80, 97D20.

Keywords. "What is Mathematics?", the image/the images of mathematics, Klein's "double discontinuity", teaching mathematics, telling stories about mathematics, the "Panorama of Mathematics" project.

1. What is Mathematics?

Defining mathematics. According to *Wikipedia* in English, in the March 2014 version, the answer to "What is Mathematics?" is

> ***Mathematics*** *is the abstract study of topics such as quantity (numbers),[2] structure,[3] space,[2] and change.[4][5][6] There is a range of views among mathematicians and philosophers as to the exact scope and definition of mathematics.[7][8]*
>
> *Mathematicians seek out patterns[9][10] and use them to formulate new conjectures. Mathematicians resolve the truth or falsity of conjectures by mathematical proof. When mathematical structures are good models of real phenomena, then mathematical reasoning can provide insight or predictions about nature. Through the use of abstraction and logic, mathematics developed from counting,*

Proceedings of the International Congress of Mathematicians, Seoul, 2014

> *calculation, measurement, and the systematic study of the shapes and motions of physical objects. Practical mathematics has been a human activity for as far back as written records exist. The research required to solve mathematical problems can take years or even centuries of sustained inquiry.*

None of this is entirely wrong, but it is also not satisfactory. Let us just point out that the fact that there is no agreement about the definition of mathematics, given as part of a definition of mathematics, puts us into logical difficulties that might have made Gödel smile.[1]

The answer given by *Wikipedia* in the current German version, reads (in our translation):

> ***Mathematics** [. . .] is a science that developed from the investigation of geometric figures and the computing with numbers. For* mathematics*, there is no commonly accepted definition; today it is usually described as a science that investigates abstract structures that it created itself by logical definitions using logic for their properties and patterns.*

This is much worse, as it portrays mathematics as a subject without any contact to, or interest from, a real world.

The borders of Mathematics. Is mathematics "stand-alone"? Could it be defined without reference to "neighboring" subjects, such as physics (which does appear in the English *Wikipedia* description)? Indeed, one possibility to characterize mathematics describes the borders/boundaries that separate it from its neighbors. Even humorous versions of such "distinguishing statements" such as

- "Mathematics is the part of physics where the experiments are cheap."
- "Mathematics is the part of philosophy where (some) statements are true — without debate or discussion."
- "Mathematics is computer science without electricity." (So "Computer science is mathematics with electricity.")

contain a lot of truth and possibly tell us a lot of "characteristics" of our subject. None of these is, of course, completely true or completely false, but they present opportunities for discussion.

What we do in Mathematics. We could also try to define mathematics by "what we do in mathematics": This is much more diverse and much more interesting than the *Wikipedia* descriptions! Could/should we describe mathematics not only as a research discipline and as a subject taught and learned at school, but also as a playground for pupils, amateurs, and professionals, as a subject that presents challenges (not only for pupils, but also for professionals as well as for amateurs), as an arena for competitions, as a source of problems, small and large, including some of the hardest problems that science has to offer, at all levels from elementary school to the millennium problems [4, 21]?

---

[1] According to *Wikipedia*, the same version, the answer to "Who is Mathematics" should be

> **Mathematics**, also known as **Allah Mathematics**, (born: **Ronald Maurice Bean**[1]) is a hip hop producer and DJ for the Wu-Tang Clan and its solo and affiliate projects.

This is not the mathematics we deal with here.

What we teach in Mathematics classes. Education bureaucrats might (and probably should) believe that the question "What is Mathematics?" is answered by high school curricula. But what answers do these give?

This takes us back to the nineteenth century controversies about what mathematics should be taught at school and at the Universities. In the German version this was a fierce debate. On the one side it saw the classical educational ideal as formulated by Wilhelm von Humboldt (who was involved in the concept for and the foundation 1806 of the Berlin University, now named Humboldt Universität, and to a certain amount shaped the modern concept of a university); here mathematics had a central role, but this was the classical "Greek" mathematics, starting from Euclid's axiomatic development of geometry, the theory of conics, and the algebra of solving polynomial equations, not only as cultural heritage, but also as a training arena for logical thinking and problem solving. On the other side of the fight were the proponents of "Realbildung": *Realgymnasien* and the technical universities that were started at that time tried to teach what was needed in commerce and industry: calculation and accounting, as well as the mathematics that could be useful for mechanical and electrical engineering — second rate education in the view of the classical German Gymnasium.

This nineteenth century debate rests on an unnatural separation into the classical, pure mathematics, and the useful, applied mathematics; a division that should have been overcome a long time ago (perhaps since the times of Archimedes), as it is unnatural as a classification tool and it is also a major obstacle to progress both in theory and in practice. Nevertheless the division into "classical" and "current" material might be useful in discussing curriculum contents — and the question for what purpose it should be taught; see our discussion in Section 8.

The Courant–Robbins answer. The title of the present paper is, of course, borrowed from the famous and very successful 1941 book by Richard Courant and Herbert Robbins [3]. However, this title is a question — what is Courant and Robbins' answer? Indeed, the book does not give an explicit definition of "What is Mathematics," but the reader is supposed to get an idea from the presentation of a diverse collection of mathematical investigations. Mathematics is much bigger and much more diverse than the picture given by the Courant–Robbins exposition. The presentation in this section was also meant to demonstrate that we need a multi-facetted picture of mathematics: One answer is not enough, we need many.

2. Why should we care?

The question "What is Mathematics?" probably does not need to be answered to motivate *why* mathematics should be taught, as long as we agree that mathematics is important.

However, a one-sided answer to the question leads to one-sided concepts of *what* mathematics should be taught.

At the same time a one-dimensional picture of "What is Mathematics" will fail to motivate kids at school to do mathematics, it will fail to motivate enough pupils to study mathematics, or even to think about mathematics studies as a possible career choice, and it will fail to motivate the right students to go into mathematics studies, or into mathematics teaching. If the answer to the question "What is Mathematics", or the implicit answer given by the public/prevailing *image* of the subject, is not attractive, then it will be very difficult to motivate *why* mathematics should be learned — and it will lead to the wrong offers and the

wrong choices as to *what* mathematics should be learned.

Indeed, would anyone consider a science that studies "abstract" structures *that it created itself* (see the German *Wikipedia* definition quoted above) interesting? Could it be relevant? If this is what mathematics is, why would or should anyone want to study this, get into this for a career? Could it be interesting and meaningful and satisfying to teach this?

Also in view of the diversity of the students' expectations and talents, we believe that one answer is plainly not enough. Some students might be motivated to learn mathematics because it is beautiful, because it is so logical, because it is sometimes surprising. Or because it is part of our cultural heritage. Others might be motivated, and not deterred, by the fact that mathematics is difficult. Others might be motivated by the fact that mathematics is useful, it is needed — in everyday life, for technology and commerce, etc. But indeed, it is not true that "the same" mathematics is needed in everyday life, for university studies, or in commerce and industry. To other students, the motivation that "it is useful" or "it is needed" will not be sufficient. All these motivations are valid, and good — and it is also totally valid and acceptable that no single one of these possible types of arguments will reach and motivate *all* these students.

Why do so many pupils and students fail in mathematics, both at school and at universities? There are certainly many reasons, but we believe that motivation is a key factor. Mathematics *is* hard. It is abstract (that is, most of it is not directly connected to everyday-life experiences). It is not considered worth-while. But a lot of the insufficent motivation comes from the fact that students and their teachers do not know "What is Mathematics."

Thus a multifacetted image of mathematics as a coherent subject, all of whose many aspects are well connected, is important for a successful teaching of mathematics to students with diverse (possible) motivations.

This leads, in turn, to two crucial aspects, to be discussed here next: What image do students have of mathematics? And then, what should teachers answer when asked "What is Mathematics"? And where and how and when could they learn that?

3. The image of Mathematics

A 2008 study by Mendick et al. [16], which was based on an extensive survey among British students, was summarized as follows:

> *Many students and undergraduates seem to think of mathematicians as old, white, middle-class men who are obsessed with their subject, lack social skills and have no personal life outside maths.*
>
> *The student's views of maths itself included narrow and inaccurate images that are often limited to numbers and basic arithmetic.*

The students' image of what mathematicians are like is very relevant and turns out to be a massive problem, as it defines possible (anti-)role models, which are crucial for any decision in the direction of "I want to be a mathematician." If the typical mathematician is viewed as an "old, white, male, middle-class nerd," then why should a gifted 16-year old girl come to think "that's what I want to be when I grow up"? Mathematics as a science, and as a profession, looses (or fails to attract) a lot of talent this way! However, this is not the topic of this presentation.

On the other hand the first and the second diagnosis of the quote from [16] belong together: The mathematicians are part of "What is Mathematics"!

And indeed, looking at the second diagnosis, if for the key word "mathematics" the *images* that spring to mind don't go beyond a *per se* meaningless "$a^2 + b^2 = c^2$" scribbled in chalk on a blackboard — then again, why should mathematics be attractive, as a subject, as a science, or as a profession?

We think that we have to look for, and work on, multi-facetted and attractive representations of mathematics by images. This could be many different, separate images, but this could also be images for "mathematics as a whole."

4. Four images for "What is Mathematics?"

Striking pictoral representations of mathematics as a whole (as well as of other sciences!) and of their change over time can be seen on the covers of the German "Was ist was" books. The history of these books starts with the series of "How and why" Wonder books published by Grosset & Dunlop, New York, since 1961, which was to present interesting subjects (starting with "Dinosaurs," "Weather," and "Electricity") to children and younger teenagers. The series was published in the US and in Great Britain in the 1960s and 1970s, but it was and is much more successful in Germany, where it was published (first in translation, then in volumes written in German) by Ragnar Tessloff since 1961. Volume 18 in the US/UK version and Volume 12 in the German version treats "Mathematics", first published in 1963 [10], but then republished with the same title but a new author and contents in 2001 [1]. While it is worthwhile to study the contents and presentation of mathematics in these volumes, we here focus on the cover illustrations (see Fig. 1), which for the German edition exist in four entirely different versions, the first one being an adaption of the original US cover of [9].

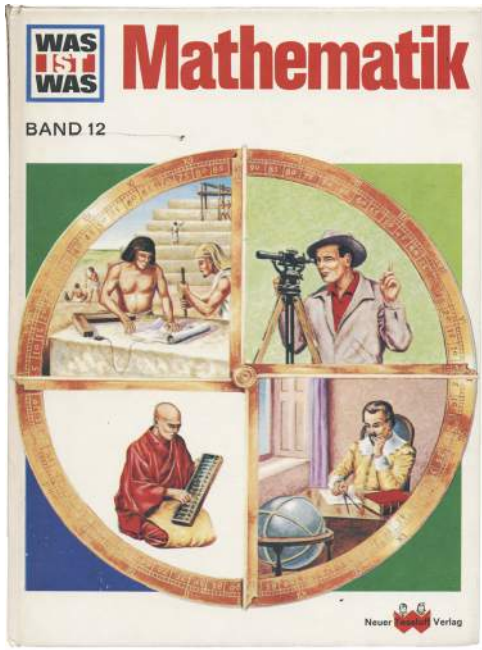
All four covers represent a view of "What is Mathematics" in a collage mode, where the first one represents mathematics as a mostly historical discipline (starting with the ancient Egyptians), while the others all contain a historical allusion (such as pyramids, Gauß, etc.) alongside with objects of mathematics (such as prime numbers or π, dices to illustrate probability, geometric shapes). One notable object is the oddly "two-colored" Möbius band on the 1983 cover, which was changed to an entirely green version in a later reprint.

One can discuss these covers with respect to their contents and their styles, and in particular in terms of attractiveness to the intended buyers/readers. What is over-emphasized? What is missing? It seems more important to us to

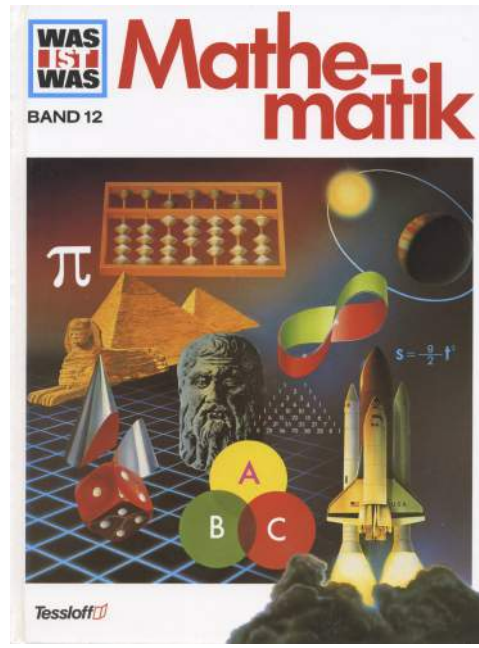
- think of our own images/representations for "What is Mathematics",
- think about how to present a multi-facetted image of "What is Mathematics" when we teach.

Indeed, the topics on the covers of the "Was ist was" volumes of course represent interesting (?) topics and items discussed in the books. But what do they add up to? We should compare this to the image of mathematics as represented by school curricula, or by the university curricula for teacher students.

In the context of mathematics images, let us mention two substantial initiatives to collect and provide images from current mathematics research, and make them available on internet platforms, thus providing fascinating, multi-facetted images of mathematics as a whole discipline:

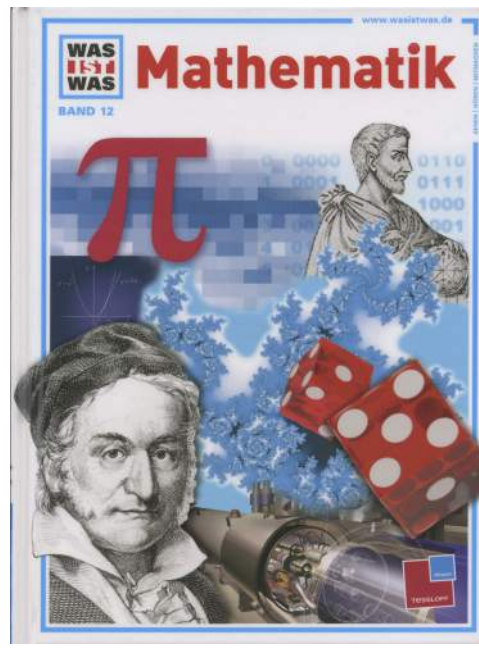


1963



1983

2001



2010

Figure 4.1. The four covers of “Was ist was. Band 12: Mathematik” [10] and [1]

- Guy Métivier et al.: "Image des Maths. La recherche mathématique en mots et en images" ["Images of Maths. Mathematical research in words and images"], CNRS, France, at images.math.cnrs.fr (texts in French)
- Andreas D. Matt, Gert-Martin Greuel et al.: "IMAGINARY. open mathematics," Mathematisches Forschungsinstitut Oberwolfach, at imaginary.org (texts in German, English, and Spanish).

The latter has developed from a highly successful travelling exhibition of mathematics images, "IMAGINARY — through the eyes of mathematics," originally created on occasion of and for the German national science year 2008 "Jahr der Mathematik. Alles was zählt" ["Year of Mathematics 2008. Everything that counts"], see www.jahr-der-mathematik.de, which was very successful in communicating a current, attractive image of mathematics to the German public — where initiatives such as the IMAGINARY exhibition had a great part in the success.

5. Teaching "What is Mathematics" to teachers

More than 100 years ago, in 1908, Felix Klein analyzed the education of teachers. In the introduction to the first volume of his "Elementary Mathematics from a Higher Standpoint" he wrote (our translation):

> *"At the beginning of his university studies, the young student is confronted with problems that do not remind him at all of what he has dealt with up to then, and of course, he forgets all these things immediately and thoroughly. When after graduation he becomes a teacher, he has to teach exactly this traditional elementary mathematics, and since he can hardly link it with his university mathematics, he soon readopts the former teaching tradition and his studies at the university become a more or less pleasant reminiscence which has no influence on his teaching."* [12]

This phenomenon — which Klein calls the *double discontinuity* — can still be observed. In effect, the teacher students "tunnel" through university: They study at university in order to get a degree, but nevertheless they afterwards teach the mathematics that they had learned in school, and possibly with the didactics they remember from their own school education. This problem observed and characterized by Klein gets even worse in a situation (which we currently observe in Germany) where there is a grave shortage of Mathematics teachers, so university students are invited to teach at high school long before graduating from university, so they have much less university education to tunnel at the time when they start to teach in school. It may also strengthen their conviction that university mathematics is not needed in order to teach.

How to avoid the double discontinuity is, of course, a major challenge for the design of university curricula for mathematics teachers. One important aspect however, is tied to the question of "What is Mathematics?": A very common highschool image/concept of mathematics, as represented by curricula, is that mathematics consists of the subjects presented by highschool curricula, that is, (elementary) geometry, algebra (in the form of arithmetic, and perhaps polynomials), plus perhaps elementary probability, calculus (differentiation and integration) in one variable — that's the mathematics highschool students get to see, so they

might think that this is all of it! Could their teachers present them a broader picture? The teachers after their highschool experience studied at university, where they probably took courses in calculus/analysis, linear algebra, classical algebra, plus some discrete mathematics, stochastics/probability, and/or numerical analysis/differential equations, perhaps a programming or "computer-oriented mathematics" course. Altogether they have seen a scope of university mathematics where no current research becomes visible, and where most of the contents is from the nineteenth century, at best. The *ideal* is, of course, that every teacher student at university has at least once experienced how "doing research on your own" feels like, but realistically this rarely happens. Indeed, teacher students would have to work and study and struggle a lot to see the fascination of mathematics on their own by doing mathematics; in reality they often do not even seriously start the tour and certainly most of them never see the "glimpse of heaven." So even if the teacher student seriously immerges into all the mathematics on the university curriculum, he/she will not get any broader image of "What is Mathematics?". Thus, even if he/she does *not* tunnel his university studies due to the double discontinuity, he/she will not come back to school with a concept that is much broader than that he/she originally gained from his/her highschool times.

Our experience is that many students (teacher students as well as classical mathematics majors) cannot name a single open problem in mathematics when graduating the university. They have no idea of what "doing mathematics" means — for example, that part of this is a struggle to find and shape the "right" concepts/definitions and in posing/developing the "right" questions and problems.

And, moreover, also the impressions and experiences from university times will get old and outdated some day: a teacher might be active at a school for several decades — while mathematics changes! Whatever is proved in mathematics does stay true, of course, and indeed standards of rigor don't change any more as much as they did in the nineteenth century, say. However, styles of proof do change (see: computer-assisted proofs, computer-checkable proofs, etc.). Also, it would be good if a teacher could name "current research focus topics": These do change over ten or twenty years. Moreover, the relevance of mathematics in "real life" has changed dramatically over the last thirty years.

6. The Panorama project

For several years, the present authors have been working on developing a course (and eventually a book [15]) called "Panorama der Mathematik" ["Panorama of Mathematics"] . It primarily addresses mathematics teacher students, and is trying to give them a panoramic view on mathematics: We try to teach an overview of the subject, how mathematics is done, who has been and is doing it, including a sketch of main developments over the last few centuries up to the present — altogether this is supposed to amount to a comprehensive (but not very detailed) outline of "What is Mathematics." This, of course, turns out to be not an easy task, since it often tends to feel like reading/teaching poetry without mastering the language. However, the approach of Panorama is complementing mathematics education in an orthogonal direction to the classic university courses, as we do not *teach* mathematics but *present* (and encourage to *explore*); according to the response we get from students they seem to feel themselves that this is valuable.

Our course has many different components and facets, which we here cast into questions about mathematics. All these questions (even the ones that "sound funny") should and can be

taken seriously, and answered as well as possible. For each of them, let us here just provide at most one line with key words for answers:

- When did mathematics start?
  Numbers and geometric figures start in stone age; the science starts with Euclid?
- How large is mathematics? How many Mathematicians are there?
  The Mathematics Genealogy Project had 178854 records as of 12 April 2014.
- How is mathematics done, what is doing research like?
  Collect (auto)biographical evidence! Recent examples: Frenkel [7], Villani [20].
- What does mathematics research do today? What are the Grand Challenges?
  The Clay Millennium problems might serve as a starting point.
- What and how many subjects and subdisciplines are there in mathematics?
  See the Mathematics Subject Classification for an overview!
- Why is there no "Mathematical Industry", as there is e.g. Chemical Industry?
  There is! See e.g. Telecommunications, Financial Industry, etc.
- What are the "key concepts" in mathematics? Do they still "drive research"?
  Numbers, shapes, dimensions, infinity, change, abstraction, . . . ; they do.
- What is mathematics "good for"?
  It is a basis for understanding the world, but also for technological progress.
- Where do we *do* mathematics in everyday life?
  Not only where we compute, but also where we read maps, plan trips, etc.
- Where do we *see* mathematics in everyday life?
  There is more maths in every smart phone than anyone learns in school.
- What are the greatest achievements of mathematics through history?
  Make your own list!

An additional question is how to make university mathematics more "sticky" for the tunneling teacher students, how to encourage or how to force them to really connect to the subject as a science. Certainly there is no single, simple, answer for this!

7. Telling stories about Mathematics

How can mathematics be made more concrete? How can we help students to connect to the subject? How can mathematics be connected to the so-called real world?

Showing applications of mathematics is a good way (and a quite beaten path). Real applications can be very difficult to *teach* since in most advanced, realistic situation a lot of different mathematical disciplines, theories and types of expertise have to come together. Nevertheless, applications give the opportunity to demonstrate the relevance and importance of mathematics. Here we want to emphasize the difference between *teaching* a topic and *telling* about it. To name a few concrete topics, the mathematics behind weather reports and climate modelling is extremely difficult and complex and advanced, but the "basic ideas" and simplified models can profitably be demonstrated in highschool, and made plausible in highschool level mathematical terms. Also success stories like the formula for the *Google* patent for *PageRank* [17], see [14], the race for the solution of larger and larger instances of

the Travelling Salesman Problem [2], or the mathematics of chip design lend themselves to "telling the story" and "showing some of the maths" at a highschool level; these are among the topics pesented in the first author's recent book [24], where he takes 24 images as the starting points for telling stories — and thus developing a broader multi-facetted picture of mathematics.

Another way to bring maths in contact with non-mathematicians is the human level. Telling stories about how maths is done and by whom is a tricky way, as can be seen from the sometimes harsh reactions on www.mathoverflow.net to postings that try to excavate the truth behind anecdotes and legends. Most mathematicians see mathematics as completely independent from the persons who explored it. History of mathematics has the tendency to become *gossip*, as Gian-Carlo Rota once put it [18]. The idea seems to be: As mathematics stands for itself, it has also to be taught that way.

This may be true for higher mathematics. However, for pupils (and therefore, also for teachers), transforming mathematicians into humans can make science more tangible, it can make research interesting as a process (and a job?), and it can be a starting/entry point for real mathematics. Therefore, stories can make mathematics more sticky. Stories cannot replace the classical approaches to teaching mathematics. But they can enhance it.

Stories are the way by which knowledge has been transferred between humans for thousands of years. (Even mathematical work can be seen as a very abstract form of storytelling from a structuralist point of view.) Why don't we try to tell more stories about mathematics, both at university and in school — not legends, not fairy tales, but meta-information on mathematics — in order to transport mathematics itself? See [23] for an attempt by the first author in this direction.

By stories, we do not only mean something like biographies, but also the way of how mathematics is created or discovered: Jack Edmonds account [6] of how he found the blossom shrink algorithm is a great story about how mathematics is actually *done*. Think of Thomas Harriot's problem about stacking cannon balls into a storage space and what Kepler made out of it: the genesis of a mathematical problem. Sometimes scientists even wrap their work into stories by their own: see e.g. Leslie Lamport's *Byzantine Generals* [13].

Telling how research is done opens another issue. At school, mathematics is traditionally taught as a closed science. Even touching open questions from research is out of question, for many good and mainly pedagogical reasons. However, this fosters the image of a perfect science where all results are available and all problems are solved — which is of course completely wrong (and moreover also a source for a faulty image of mathematics among undergraduates).

Of course, working with open questions in school is a difficult task. None of the big open questions can be solved with an elementary mathematical toolbox; many of them are not even accessible as questions. So the big fear of discouraging pupils is well justified. On the other hand, why not explore mathematics by showing how questions often pop up on the way? Posing questions in and about mathematics could lead to interesting answers — in particular to the question of "What is Mathematics, Really?"

8. Three times Mathematics at school?

So, what is mathematics? With school education in mind, the first author has argued in [22] that we are trying cover three aspects the same time, which one should consider separately

and to a certain extent also teach separately:

Mathematics I: A collection of basic tools, part of everyone's survival kit for modern-day life — this includes everything, but actually not much more than, what was covered by Adam Ries' "Rechenbüchlein" ["Little Book on Computing"] first published in 1522, nearly 500 years ago;

Mathematics II: A field of knowledge with a long history, which is a part of our culture and an art, but also a very productive basis (indeed a production factor) for all modern key technologies. This is a "'story-telling" subject.

Mathematics III: An introduction to mathematics as a science — an important, highly developed, active, huge research field.

Looking at current highschool instruction, there is still a huge emphasis on Mathematics I, with a rather mechanical instruction on arithmetic, "how to compute correctly," and basic problem solving, plus a rather formal way of teaching Mathematics III as a preparation for possible university studies in mathematics, sciences or engineering. Mathematics II, which should provide a major component of teaching "What is Mathematics," is largely missing. However, this part also could and must provide motivation for studying Mathematics I or III!

9. What is Mathematics, really?

There are many, and many different, valid answers to the Courant–Robbins question "What is Mathematics?"

A more philosophical one is given by Reuben Hersh's book "What is Mathematics, Really?" [11], and there are more psychological ones, on the working level. Classics include Jacques Hadamard's "Essay on the Psychology of Invention in the Mathematical Field" and Henri Poincaré's essays on methodology; a more recent approach is Devlin's "Introduction to Mathematical Thinking" [5], or Villani's book [20].

And there have been many attempts to describe mathematics in encyclopedic form over the last few centuries. Probably the most recent one is the gargantuan "Princeton Companion to Mathematics" [8], edited by Tim Gowers et al., which indeed is a "Princeton Companion to Pure Mathematics."

However, at a time where *Zentralblatt MATH* counts more than 100.000 papers and books per year, and 24821 submissions to the math and math-ph sections of arXiv.org in 2013, it is hopeless to give a compact and simple description of what mathematics really is, even if we had only the "current research discipline" in mind. The discussions about the classification of mathematics show how difficult it is to cut the science into slices, and it is even debatable whether there is any meaningful way to separate applied research from pure mathematics.

Probably the most diplomatic way is to acknowledge that there are "many mathematics." Some years ago Tao [19] gave an open list of mathematics that is/are good for different purposes — from "problem-solving mathematics" and "useful mathematics" to "definitive mathematics", and wrote:

> *"As the above list demonstrates, the concept of mathematical quality is a high-dimensional one, and lacks an obvious canonical total ordering. I believe this*

is because mathematics is itself complex and high-dimensional, and evolves in unexpected and adaptive ways; each of the above qualities represents a different way in which we as a community improve our understanding and usage of the subject."

In this sense, many answers to "What is Mathematics?" probably show as much about the persons who give the answers as they manage to characterize the subject.

Acknowledgements. The authors' work has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 247029, the DFG Research Center MATHEON, and the DFG Collaborative Research Center TRR 109 "Discretization in Geometry and Dynamics."

References

[1] Wolfgang Blum, *Was ist was. Band 12: Mathematik*, Tessloff Verlag, Nürnberg, 2001. Revised version, with new cover, 2010.

[2] William Cook, *In Pursuit of the Traveling Salesman: Mathematics at the Limits of Computation*, Princeton University Press, Princeton NJ, 2011.

[3] Richard Courant and Herbert Robbins, *What Is Mathematics? An Elementary Approach to Ideas and Methods*, Oxford University Press, 1941. Second edition, edited by I. Stewart, 1996.

[4] George Csicsery, *Hard Problems. the Road to the World's Toughest Math Contest*, Documentary film, 82 minutes (feature)/45 minutes (classroom version), Mathematical Association of America, Washington DC, 2008.

[5] Keith J. Devlin, *Introduction to Mathematical Thinking*, published by Keith Devlin, Palo Alto CA, 2012.

[6] Jack Edmonds, *A glimpse of heaven*, in: J. K. Lenstra, A. Schrijver, and A. Rinnooy Kan, eds., "History of Mathematical Programming — A Collection of Personal Reminiscences," CWI and North-Holland, Amsterdam, 1991, pp. 32–54.

[7] Edward Frenkel, *Love & Math. The Heart of Hidden Reality*, Basic Books/Perseus Books, Philadelphia PA, 2013.

[8] Timothy Gowers, Imre Leader, and June Barrow-Green, eds., *The Princeton Companion to Mathematics*, Princeton University Press, Princeton NJ, 2008.

[9] Esther Harris Highland and Harold Joseph Highland, *The How and Why Wonder Book of Mathematics*, Grosset & Dunlop, New York, 1961.

[10] _____, *Was ist was. Band 12: Mathematik*, Neuer Tessloff Verlag, Hamburg, 1963. Revised edition 1969. New cover 1983.

[11] Reuben Hersh, *What is Mathematics, Really?* Oxford University Press, Oxford, 1997.

[12] Felix Klein, *Elementarmathematik vom höheren Standpunkte aus. Teil I: Arithmetik, Algebra, Analysis*, B. G. Teubner, Leipzig, 1908. Vierte Auflage: Springer, Heidelberg 1933.

[13] Leslie Lamport, Robert Shostak, and Marshall Pease, *The Byzantine Generals Problem*, ACM Transactions on Programming Languages and Systems **4** (1982), 382–401.

[14] Amy N. Langville and Carl D. Meyer, *Google's PageRank and Beyond. The Science of Search Engine Rankings*, Princeton University Press, Princeton and Oxford, 2006.

[15] Andreas Loos and Günter M. Ziegler, *Panorama der Mathematik*, Springer Spektrum, Heidelberg, 2015, in preparation.

[16] Heather Mendick, Debbie Epstein, and Marie-Pierre Moreau, *Mathematical images and identities: Education, entertainment, social justice*, Institute for Policy Studies in Education, London Metropolitan University, London 2008.

[17] Lawrence Page, *Method for node ranking in a linked database*, United States Patent No. US 6,285,999 B1, 4. September 2001 (submitted: 9. Januar 1998), http://www.google.com/patents/US6285999.

[18] Gian-Carlo Rota, *Indiscrete Thoughts*, Birkhäuser, Basel, 1996.

[19] Terence Tao, *What is good Mathematics?*, Bulletin Amer. Math. Soc. (4)**44** (2007), 623–634.

[20] Cédric Villani, *Théorème vivant*, Bernard Grasset, Paris, 2012 (in French).

[21] Günter M. Ziegler, *Three competitions*, in: D. Schleicher and M. Lackmann, eds., "Invitation to Mathematics. From Competition to Research," Springer, Berlin Heidelberg, 2011, pp. 195–205.

[22] _____, *Mathematics school education provides answers — to which questions?*, EMS Newsletter No. 84, June 2012, 8–11.

[23] _____, *Do I Count? Stories from Mathematics*, CRC Press/Taylor & Francis, Boca Raton FL, 2013. English translation of "Darf ich Zahlen? Geschichten aus der Mathematik", Piper, München, 2010.

[24] _____, *Mathematik — Das ist doch keine Kunst!* Knaus, München, 2013.

Inst. Mathematics, Freie Universität Berlin, Arnimallee 2, D-14195 Berlin, Germany
E-mail: ziegler@math.fu-berlin.de

Inst. Mathematics, Freie Universität Berlin, Arnimallee 7, D-14195 Berlin, Germany
E-mail: loos@math.fu-berlin.de

19. History of Mathematics

Knowledge and power: A social history of the transmission of mathematics between China and Europe during the Kangxi reign (1662–1722)

Qi Han

Abstract. In the last few decades much research has been devoted to the interaction of European and Chinese mathematics in the seventeenth and eighteenth centuries. Scholars have begun to consider social and political factors in their studies of Chinese mathematics. This approach, however desirable, needs more systematic exploration. Drawing on research findings in social and political history, I will analyse why the Kangxi Emperor (1654-1722) began to be interested in European mathematics and how he used his newly acquired mathematical knowledge as a tool to control and impress Chinese official scholars and so consolidate his power. In addition, I will point out the reasons why he changed his attitude toward Western learning and established an Academy of Mathematics in 1713. Then I explore how European mathematical books were introduced and circulated in the Kangxi reign (1662-1722). Further I discuss why the Kangxi Emperor became interested in traditional Chinese mathematics. Finally, using both Chinese and European sources, I discuss the study of *The Book of Changes* (易经) at the imperial court and its link to the French Jesuit Joachim Bouvet (1656-1730) and the German philosopher Leibniz.

Mathematics Subject Classification (2010). 01A25, 01A45, 01A50.

Keywords. Chinese mathematics, Jesuits, Kangxi Emperor, Leibniz, transmission, 17th century, 18th century.

After the Italian Jesuit Matteo Ricci (1552-1610) arrived in China in 1582, Chinese science entered a new era. The aims of the Jesuits in China were, of course, primarily missionary, and from the beginning they used science merely as a means of arousing Chinese scholars' interest in Christianity. However, their influence in China was to prove effective mostly in the field of science itself. In the late seventeenth and early eighteenth centuries, they played a leading role in the transmission of mathematical knowledge between China and Europe.

The Kangxi Emperor (1654-1722), the second ruler of the Qing dynasty (1644-1911), reigned over this vast empire from 1662 to 1722. As a Manchu monarch, he had from childhood followed the Manchu traditions of archery and horse-riding. At the same time he received a good education in the traditional Confucian classics from his high officials. And, he played an essential role in the transmission of Western mathematics to China.

In the 1660s, greatly impressed by a controversy between Jesuit missionaries and Chinese scholars on scientific and religious matters, he began to study European mathematics seriously. In the long history of China such an interest was unusual for an emperor. In this talk I

Proceedings of the International Congress of Mathematicians, Seoul, 2014

would like to begin by considering why the Kangxi Emperor started to study Western mathematics and how he used it as a means to show off in front of Chinese officials, particularly in a celebrated episode that happened in the Forbidden City in 1692. In addition, by examining the emperor's attitude toward Western science, I will analyze why in 1713 he launched a new Academy of Mathematics (算学馆). Then I will talk about the circulation and translation of European mathematical books in China and explore why the emperor became interested in traditional Chinese mathematics. Finally I will talk about the leading role the French Jesuit Joachim Bouvet (1656-1730) played in bridging the shared interests of the German philosopher Leibniz and the Kangxi Emperor in their study of *The Book of Changes* (易经) and a binary system of numbers.

1. The Kangxi Emperor's interest in European mathematics

Shortly after the establishment of the Qing empire in 1644, its first Manchu emperor Shunzhi (1638-1661), invited the German Jesuit Adam Schall von Bell (1592-1666) to be director of the Imperial Board of Astronomy (钦天监). At Schall's suggestion, Shunzhi had the Qing adopt the European astronomical system. However, after the Kangxi Emperor came to power in 1662, this consensus came under attack. In 1664, a conservative Chinese scholar Yang Guangxian (杨光先, 1597-1669) launched a "calendar case," that harshly criticized European astronomy. Schall and his colleagues were arrested, and several Chinese Christian astronomers put to death. This event deeply impressed the youthful Kangxi Emperor, as he later recounted to his sons:

> *You only know that I'm versed in mathematics. But you don't know the reason why I study mathematics. When I was young, the Chinese officials and the Westerners at the Imperial Board of Astronomy were on unfriendly terms with each other. They made accusations against one another. It almost came to capital punishment. Yang Guangxian and Adam Schall von Bell (actually Ferdinand Verbiest, as Kangxi's memory here was faulty) competed in measuring the length of the sun's shadow in front of nine chief ministers at Wu Men Gate. Unfortunately, among those ministers there was no one who knew about these methods. I realized that if I didn't know it myself, how could I judge true from false? So I very eagerly determined to study mathematics. (Kangxi Emperor's Instructions to His Children* 庭训格言*,Yongzheng edition, pp.78-79)*

Thus, it was this calendar dispute that prompted the Kangxi Emperor to study Western mathematics. From his youth, he became very interested in Western science and invited the Belgian Jesuit Ferdinand Verbiest (1623-1688) to be his earliest scientific tutor. At that time, Verbiest was worried about the severe shortage of Jesuit mathematicians in China, and in 1678 he wrote a letter urgently appealing to Jesuits all over Europe to join the China mission. Because he was in his late fifties, the emperor soon felt that it was urgent to find a person to replace him at the Imperial Board of Astronomy. Therefore, Verbiest recommended Antoine Thomas (1644-1709), a Jesuit from Namur who was well-versed in mathematics and had been stationed in Macao since 1682. At the emperor's request, Thomas was summoned to Beijing in 1685.

Thomas was a crucial figure in the history of mathematics during the Kangxi court. Before leaving for China, he had written *Synopsis mathematica* (Douai, 1685), a book for be-

ginners and for the Jesuits in China anxious to propagate the gospel through their knowledge of mathematics and astronomy. After his arrival in Beijing in 1685, Thomas was given the honour of being a tutor to the emperor ([25, 11]).

In response to Verbiest's urgent appeal of 1678, Jean de Fontaney (1643-1710), Joachim Bouvet, J.-F. Gerbillon (1654-1707) and other two Jesuits came to Beijing. Sent by Louis XIV as "the King's Mathematicians", they were expected to glorify the French King, propagate Christian doctrines, benefit science and arts, and thereby reduce Portuguese sea power in East Asia. They were also charged with making astronomical observations, investigating native Chinese flora and fauna, and learning other technical arts.

After their arrival in Beijing in 1688, the Kangxi Emperor consulted these French mathematicians on European arts and science, and systematically set about studying these aspects of Western learning. From 1689 to 1691 Bouvet, Gerbillon, Thomas, and T. Pereira (1645-1708) frequently taught mathematics to the emperor ([23]). They translated mathematical books into Chinese and Manchu including the *Elemens de Geometrie* by the French Jesuit I.-G. Pardies (1636-1673) and *Synopsis Mathematica*. They also introduced to the imperial court many mathematical instruments (e.g., calculating machines, Napier's rods, proportional compasses, surveying instruments, etc.), many of which are preserved today in the Palace Museum in Beijing. To help the emperor in his study of European mathematics, these Jesuits designed a special mathematical table and models for teaching solid geometry.

2. Mathematics as a tool to control and impress Chinese official scholars

Ample evidence shows that during the early Qing dynasty scholars of Han Chinese origin perceived the Manchu rulers as a comparatively uncivilized ethnic minority; that is, they were both alien and barbarian. Even after the full "reunification" of the empire under Manchu rule in 1683, there were numerous political and cultural conflicts between Manchus and Han Chinese. In order to promote Manchu prestige, the Kangxi Emperor studied not only the traditional Chinese classics, but also Western learning, even attempting to show his Chinese subjects that his command of Western learning was superior to theirs.

The ongoing tension between the Manchus and Han Chinese was a great concern to the emperor, leading to his own recurrent distrust of Chinese officials. To test their honesty and loyalty, examine their knowledge, and reduce their sycophancy, he commonly used his newly acquired Western learning. In 1689 the emperor visited the observatory in Nanjing during his southern tour. He had just completed several months of ardent study of European mathematics, and the questions he posed to his officials demonstrated not only his superior command of mathematics but also his wish to shape the nature of their learning ([5, 6]).

In addition, the Kangxi Emperor seized the opportunity to show off his mathematics ability at the court. On February 20, 1692, Kangxi summoned his high ministers and even a Chinese scholar versed in mathematics to the Hall of Heavenly Purity (乾清门) for a lecture on the relationship between music and mathematics, as well as on the ratio of the circumference of a circle to its diameter. He also ordered the installation of a gnomon, and personally drew a line to demonstrate his understanding of how a gnomon functioned. He predicted the location where the shadow would reach at high noon and ordered his ministers to observe the shadow. At exactly high noon, the length of the shadow just touched the line drawn by the emperor, missing it by not so much as a hair. This event is recounted in not just the court's records but also in the private records of ministers assembled to view this spectacle of impe-

rial sagacity. For instance, Wang Xi (王熙, 1628-1703), the Minister of Rites, tells of being summoned on that day, along with other Manchu and Han ministers as well as scholars of the Hanlin Academy, to the Hall of Heavenly Purity for imperial instruction. The emperor ordered them to calculate—using methods in the classic Chinese book of mathematics (i.e., *The nine chapters of mathematical procedures*)—and to observe the movement of the sun. In the space of half a day, a number of musical, mathematical, and astronomical calculations were made, and a method for calculating the volume of moving water was demonstrated. Kangxi's ability left a lasting impression on the assembled officials ([19]): "They received humbly the emperor's lessons, heard what has never before been heard, saw what was never before seen, and [their] joy knew no limits." (圣祖实录, juan 154) In addition, they also felt an undefined sense of inadequacy: "After the meeting, we were joyful yet deeply ashamed of the shallowness of our knowledge. We had stubbornly held onto stale knowledge and were unknowingly seduced by it." They then proposed to Kangxi plans to compile new books on music, mathematics and astronomy, and so "preserve this knowledge forever."(Zhang Yushu 张玉书:*Collected Writings of Zhang Yushu* 张文贞公集 [1792 ed.], 2, pp.9-11.) However, nothing of that kind happened before 1713.

Kangxi's actions also had a political motive, as the lecture and demonstration provided him an opportunity to show his "genius" in front of his ministers and challenge Chinese officials' presumptions about the superiority of their cultural and mathematical learning. His success in this court session of 1692 had resulted from more than two years of systematic study. The Jesuits had given Kangxi access to "new knowledge" from Europe, and thereby provided him with the basis for his memorable "performance." On several other occasions, Kangxi also used his newly acquired skills to put the Confucian elite in their place, even in 1702 publicly criticizing Chinese scholars as completely ignorant of mathematics.

During the course of his education, the emperor learned many things from the Jesuits, and became quite knowledgeable in scientific matters. As soon as he received any new information from his Jesuit tutors, he tried to teach what he understood and manifest his scientific ability in front of his ministers. He frequently asked his subjects mathematical questions that he had just learned from the Jesuits. Sometimes, he even personally instructed young scholars in mathematics, and displayed his wide knowledge in front of his ministers during field trips that involved map-surveying or visiting river dam construction projects. Through many such demonstrations, the emperor won their admiration for his talent in mathematics. Indeed, he seemed very pleased by the ministerial applause that greeted his demonstrations of scientific ability and erudite knowledge. As a Manchu ruler, he wanted to exhibit to Chinese the cultural and intellectual accomplishments that he had and they did not, to confirm why he and his Manchu family were the Sons of Heaven ruling China and its people.

3. The political background to the launching of a new project

Sent to China as the Papal Legate to solve the problems of the Rites Controversy, Carlo Tommaso Maillard de Tournon (1668-1710) arrived in Beijing in 1705, at a time when disagreements between the Qing court and the Catholic church had become serious. The Legate wanted to prohibit the Chinese Christians of ancestral worship and other traditional Chinese practices and, in February 1707, issued the decree of Nanjing which condemned the practices in question as superstitious. This decree offended the Kangxi Emperor greatly and fed his distrust of the Europeans. In late 1706, after listening to a lecture on Zhu Xi's (朱熹) learning,

the Kangxi Emperor summoned two of his ministers and said to them: "Do you know that the Westerners are increasingly mischievous? They're even attacking Confucius! The reason why I treat them well is merely to make use of their skills and arts." (Li Guangdi, *Rongcun yulu xuji* 榕村语录续集, juan 6)

De Tournon's legation in fact is a turning point in the imperial court's relationship to the Catholic church. Kangxi, once so friendly with the missionaries, became suspicious of and even hostile to them, a change of mind that is evident in many official documents and memorials and that quickly became known throughout the empire.

His stance was not purely emotional or political. In 1706, the emperor had believed that European methods of predicting astronomical phenomena were much more precise. Yet in 1711 in observing the shadows of the sun at the summer solstice, he found that the calculation of the Imperial Board of Astronomy, based on European methods, was not accurate and so he changed his conviction. ([19]) He realized that the European astronomical system—at least the one he had access to—was out-of-date, and a new compendium of astronomy and mathematics must be compiled.

For this purpose, the emperor issued an edict in 1713 to establish an Academy of Mathematics at the Studio for the Cultivation of the Youth (蒙养斋). Located in an imperial villa in the northwest suburbs of Beijing, it was to be the Emperor's academy. Although four Jesuits—P. Jartoux (1669-1720), J.-F. Foucquet (1665-1741), F. Thilisch (1670-1715) and K. Slavicek (1678-1735)—were eventually asked to teach there, it operated largely independent of the Westerners at his court ([10]). After interviewing more than three hundred candidates, he personally recruited seventy-two young Chinese, Manchu and Mongolian scholars well versed in mathematics to serve in this academy ([14], [15]). The compilation of a compendium of astronomical, mathematical and musical texts, entitled *The Origin of pitchpipes and calendar* (律历渊源), was one of its main goals. This imperially composed work comprises three parts:*Basic Mathematical Principles* (数理精蕴), *Imperially composed calendar* (钦若历书, later called *Compendium of the calendar*, 历象考成), and the *Exact meaning of pitchpipes* (律吕正义). In addition, many astronomical observations were made by members of this academy in order to collect data to be used for various problems, most notably, the obliquity of the ecliptic. Based on their observations, Chinese mathematicians corrected some astronomical data and completed the compilation of the *Imperially composed calendar* (printed in about 1722).

After the Academy of Mathematics was established, the Kangxi Emperor worked as a mathematics tutor, at times even teaching some of the young students. Mei Juecheng (梅毂成, 1681-1763), who was taught by the emperor, wrote in his book titled *Pearls Lost in the Red River* (赤水遗珍, 1745):

> *Later I served at the imperial court. I am indebted to the benevolent Emperor (Kangxi) for teaching me the 'method of borrowing radix and powers (借根方)'. And His Majesty instructed, 'The Westerners name this book as a-er-re-ba-da (i.e. algebra), which means the method from the East (东来法).' I read it respectfully. This method, mysterious and wonderful, is really the guide to mathematical methods.*

A Manchu memorial also mentioned that Kangxi taught geometry to the students. The emperor's teaching was largely symbolic, and should be seen as part of his attempt to display his scientific abilities in front of Han Chinese officials.

In addition, Kangxi strove to assure that his children received a good education in Western

science, to help them run the empire. His third son Yinzhi (胤祉, 1677-1732), knowledgeable of Western science thanks to his studies with Antoine Thomas, was entrusted with supervising the Academy of Mathematics. Of the other princes who learned a great deal from the Jesuits, the thirteenth and the sixteenth sons proved proficient in the Western sciences and played a leading role in the organization of scientific activities ([9]).

4. The circulation and translation of mathematical books

In order to understand better the circulation through the Chinese empire of the learning and knowledge brought from Europe, it is essential to observe the different ways European mathematical books arrived in China and were kept in three Catholic churches in Beijing.

In the seventeenth and early eighteenth centuries, when the Jesuits and Chinese were translating books of European mathematics, they sometimes mentioned the original source texts. However in most cases, European sources were not mentioned. Since most of the books compiled at the Kangxi court were imperially commissioned, the names of European authors and the titles of their source texts were seldom cited. This practice has led to great difficulty in identifying the origins of the mathematics in these Chinese-language books. Yet, like the accounts so far written of Jesuit libraries in China, this practice of not citing Western sources has also downplayed the amount of European mathematical learning that was transmitted to China at this time.

Since the French Jesuits like Bouvet, Jartoux and Foucquet ([28, 20]) served as court mathematicians, they played a leading role in introducing a number of European mathematical books into Chinese. Their mathematical collections should be seriously considered.

Recently scholars have begun to do excellent research on the circulation of European books and the Jesuit libraries in different places in China, helping us understand how and why some mathematical books were introduced and translated. According to Noël Golvers' research, many European mathematics books were brought to China and preserved in Jesuit libraries there, such as *Elémens de géometrie* (I.-G. Pardies), *Nouveaux élemens de géometrie* (A. Arnauld), *Cursus seu mundus mathematicus* (C. F. Dechales), *Nouveaux elémens des mathématiques* (J. Prestet), *Recueil des traités de mathématiques* (P. Hoste), *Nouveaux élemens d'arithmétique et d'algebre, ou introduction aux mathématiques* (T. F. de Lagny), *Elémens des mathématiques* (B. Lamy), *Récréations mathématiques* (J. Ozanam), *Traité d'algebre* (M. Rolle), etc. In addition to these elementary mathematical books, more advanced books like *Analyse des infiniment petits* (Marquis de l'Hospital),*Opera mathematica* (J. Wallis), and *Méthode pour la mesure des surfaces* (L. Carré) were also on the shelves of these Jesuit libraries. Many of these books were collected and used by Foucquet ([3], pp.190-194).

In addition to these mathematical books, numerous journals of European institutions like *Histoire de l'Académie Royale des Sciences, Journal des sçavans, Philosophical Transactions*, and *Acta Eruditorum* arrived in Beijing. The circulation of these books and journals will help to understand better the original sources used during the compilation of mathematical encyclopaedia like the *Basic Mathematical Principles*.

In fact, I have checked carefully the mathematical sources used to compile the *Basic Mathematical Principles*, and I found that the new mathematical knowledge introduced by the Jesuits in this book included methods for calculating logarithms based on the *Arithmetica Logarithmica* (1624) of the English mathematician Henry Briggs (1561-1630), the logarithm

table of the Dutch mathematician Adriaan Vlacq (1600-1667), *Trigonometria artificialis: sive magnvs canon triangvlorvm logarithmicvs* (Govdae, 1633), *Elémens de géometrie* of Pardies, *Synopsis mathematica* of Antoine Thomas, and the method for solving equations of higher degree ([4], [13]). In addition, three formulae of infinite series expansion were introduced by Jartoux, and subsequently copied by Mei Juecheng in his *Pearls Lost in the Red River*. Jartoux was skilled at geometry and analysis, was also familiar with the progress of calculus in Europe as shown in his correspondence with Leibniz, and so was more than capable of introducing advanced mathematical knowledge to Qing mathematicians among whom were Ming Antu (明安图, ?-1763?), a Mongolian mathematician, and a Lama from Tibet.

Furthermore, many handwritten or printed mathematical tables, including those of logarithms, sine, cosine, tangents and trigonometric logarithms, are still preserved today in the Palace Museum in Beijing. Some of them were copied from the mathematical tables of Jacques Ozanam and frequently used by the Kangxi emperor on his field trips outside of Beijing ([4]).

In other words, recent research has demonstrated that during the Kangxi era much more European mathematical knowledge circulated in imperially edited or commissioned works than previously thought, in part due to inadequate cross-referencing. Noël Golvers' work has also shown the existence in Beijing and some other Chinese cities of Western books of mathematics and other sciences, and in all likelihood in the near future we can expect him to detail how some of this learning left the sanctuary of Jesuit libraries in Beijing for wider circulation in China during the seventeenth and eighteenth centuries.

5. Chinese vs Western?: The Kangxi Emperor's interest in traditional mathematics

After 1669, the Kangxi Emperor believed that Western science was far superior to Chinese science and often expressed this opinion in conversations with his high officials. However, shortly after beginning his ardent study of mathematics, he proposed the theory of "The Chinese Origin of Western Learning"(西学中源) in 1703, an idea that the Western mathematical sciences had originated in China. In his short essay entitled "Imperially composed treatise on the derivation of triangles" (御制三角形推算法论), the Kangxi Emperor first explained why he had studied Western science and then talked about the relationship between the European and Chinese traditional calendars:

> *Some believe that the ancient and modern methods are different. Actually they do not know the calendar deeply. The calendar, which originated in China, was transmitted to the Far West. The Westerners kept it, made observations endlessly and revised it every year. Therefore, their calendar is quite precise. (Kangxi, Yuzhi wenji* 御制文集*, Ji 3, juan 19.* [5]*)*

The reason why he proposed this theory of "The Chinese Origin of Western Learning" is quite interesting. As a Manchu ruler who governed the Han Chinese, he intended to embrace Confucianism. If he learned Western mathematics, he therefore might be regarded as alien by Chinese officials. To some extent, his aim in advocating this theory was to mitigate such criticism, to play down any dispute between Western and Chinese learning and to dull the anti-foreign antagonisms in the anti-Christianity movement. At the same time, it also

provided Han Chinese with an excellent justification to learn European mathematics and astronomy. Having developed from ancient Chinese sciences, this learning could be judged as fundamentally Chinese. Thanks to the emperor's advocacy and subsequent propagation by Chinese scholars, this theory became widely known and influenced the study of the mathematical sciences in China during the 18th and 19th centuries.

In order to strengthen the impression that he also was a master of traditional Confucian learning, the Kangxi Emperor paid attention to traditional Chinese mathematical works, such as those written by Zhu Zaiyu (朱载堉, 1536-1611) and Cheng Dawei (程大位, 1533-1606).

As a "King's Mathematician," Bouvet taught European mathematics to the emperor. But, as is shown by his study of *The Book of Changes*, he also studied Chinese mathematics. In a memorial of about 1711 Wang Daohua (王道化), an official at the Imperial Household, mentioned that Bouvet was studying the magic squares in the *Systematic Treatise of Mathematical Calculation* (算法统宗, 1593), a treatise written by the Ming mathematician Cheng Dawei. Probably the reason why the emperor was interested in the *Systematic Treatise of Mathematical Calculation* is that Bouvet's study helped him to understand mathematical knowledge contained in traditional Chinese mathematics, like that of Pascal's triangle. On 14 August 1712, the emperor issued an edict, in which he declared that the *Systematic Treatise of Mathematical Calculation* was very useful and ordered officials at the Imperial Household to search for it and present it to him. Three days later, they submitted it to him. Having read it, Kangxi said that this book was very good. When the news spread about the emperor's interest in the *Systematic Treatise of Mathematical Calculation*, a new edition was privately printed. In his preface to the new edition, Cheng Shisui (程世绥), a grandson of Cheng Dawei, writes:

> *When I arrived in the capital, the Son of Heaven was interested in musical and calendrical sciences. He had established a mathematics school and a bureau to compile mathematical books. Many erudite scholars from various parts of the empire assembled in crowds in the capital. When I had the leisure, I followed them. They earnestly and approvingly spoke of the Systematic Treatise of Mathematical Calculation. They thought that this book really epitomized the mathematical sciences and that it had been highly praised by the emperor. Famous scholars and high officials also competed to buy it as something precious. Therefore I came back and read it very carefully. I found that the book was very well organized... The classic of the Nine chapters [of mathematical procedures] and the method of multiplication and division are all abundantly clear. (Cheng Dawei, Zhizhi Suanfa tongzong* 直指算法统宗*, 1716 ed.* [7]*)*

In addition, *The Gnomon of Zhou [Dynasty]* (周髀算经), the oldest classic of ancient Chinese astronomy, aroused the emperor's interest. In 1711, Kangxi's third son Yinzhi mentioned that the emperor had read the *The Gnomon of Zhou [Dynasty]* and made some commentaries on it. The emperor's interest also influenced the imperial compilers of the *Basic Mathematical Principles*. At the front of its opening chapter, the compilers inserted the *Annotation of The Gnomon of Zhou [Dynasty]* (周髀经解), thereby underlining their message that China had the most ancient tradition in mathematics and thus was the fount of mathematical knowledge.

6. Between Leibniz and Kangxi: Bouvet's study of *The Book of Changes* and binary system

In addition to the *The Gnomon of Zhou [Dynasty]*, the Kangxi Emperor was interested in the Confucian classic *The Book of Changes*. His ardent interest in this book derived in large part from his relationship with Bouvet, whose own interest in *The Book of Changes* had been inspired by the German philosopher Leibniz (1646-1716).

Leibniz had been interested in China from no later than 1666. In 1689, he began to correspond with the Italian Jesuit C. F. Grimaldi (1638-1712) after they had met in Rome. On 19 July 1689, Leibniz asked Grimaldi a couple of questions, in which he wondered if there had been geometrical demonstrations in ancient China. ([27], p.5) He also asked Bouvet similar questions ([2], Introduction, p.3) and in 1697 published the *Latest News from China (Novissima Sinica)*. On October 18, having read this book after his return to France, Bouvet wrote to Leibniz at Fontainbleau. From then on, they kept in close contact.

In his letter dated 15 February 1701, Leibniz first introduced the idea of the binary system to Bouvet. From a theological point of view, he believed that all numbers can be derived from 1 and 0. He thought that his study of the binary system would have a great impact on Chinese philosophers and even interest the Kangxi Emperor. Hence, he strongly encouraged Bouvet to present it to the emperor. ([27], p.139)

On 4 November 1701, in a long letter addressed to Leibniz Bouvet enclosed the diagrams of Fuxi (伏羲) in *The Book of Changes*. Bouvet mentioned his questions about the binary system and claimed that some Chinese records were identical with Leibniz's mathematical ideas. He suggested that Leibniz use hexagrams (六爻, *liuyao*) to explain the binary system and thought that the trigrams (八卦, *bagua*) of Fuxi were the origin of mathematical wisdom. He also pointed out that Leibniz's numerical table was exactly the same as that which Fuxi had used to form his system. Though Bouvet believed that it thus would not be regarded as a new science, at least in China, he was convinced that Leibniz's study opened a new route for people to understand the real system of nature. ([27], p.150)

Although Leibniz had been working on a binary system for a long time, he had not planned to publish it. The year 1700 was important because Leibniz became a corresponding member of the Royal Academy of Sciences in Paris. On 1 April 1703 Bouvet's letter of 4 November 1701 reached Leibniz in Berlin. Within a week of receiving it, Leibniz communicated the discovery to his friend Carlo Mauritio Vota, the confessor of the King of Poland, and sent it onto the Abbé Bignon for publication in the journal of the Paris Academy. ([1], pp.245-247) This paper, entitled "Explication de l'Arithmetique binaire", was published in the *Histoire de l'Académie Royale des Sciences* (Année MDCCIII) in 1705. And so, we can conclude that thanks to Bouvet's letter Leibniz was stimulated to publish his paper on the binary system.

As mentioned above, Leibniz had suggested to Bouvet that he submit Leibniz's idea on the binary system to the Kangxi Emperor. Interestingly enough, the 1705 edition of the *Histoire de l'Académie Royale des Sciences* with this article was presented to the emperor in 1714 when Bouvet, Kilian Stumpf (1655-1720), and other Jesuits were summoned to the imperial court. The Kangxi Emperor thereby got to know of Leibniz's name and became curious about his mathematics, asking the Jesuits to tell him as soon as possible what was worth knowing of it. The archival documents vividly record the ensuing dialogue among the Jesuits and the emperor ([17]), but as Stumpf did not recognize the importance of Leibniz's paper, he did not have it translated for the emperor. Hence, Leibniz's binary system would have to wait until the twentieth century to be introduced to China.

Perhaps after his correspondence with Leibniz, Bouvet was stimulated to study *The Book of Changes* seriously for its mathematics. His research even aroused the Kangxi Emperor's interest. In 1711, the Emperor put forward the new idea that mathematical principles all derived from *The Book of Changes* and in the following year claimed that Western methods were identical with the numerical principles in this book. It is interesting that Kangxi's view was influenced by Bouvet, who served as a court scholar in the compilation of *The Book of Changes*. While it was not Bouvet's purpose to prove this theory, he diligently investigated the Chinese classics from a Figurist point of view. ([26]) However, his results unintentionally provided support for the emperor's theory of "The Chinese Origin of Western Learning."

7. Concluding remarks

The Kangxi period was very crucial in the transmission of European mathematical science to China. Kangxi's interest in mathematics was widely known in the empire, and in order to win the emperor's favour, some scholars began to study mathematics and train young mathematicians. The importance of mathematics was recognized by Chinese literati; this recognition in turn stimulated the development of mathematics in China.

However, in the end the spread of Western science during the Kangxi reign did not succeed. Kangxi tried to associate his scientific studies with his political interests, by using his knowledge of science as an aid in statecraft to control Chinese officials and even the missionaries. Early on in his reign, relations between the Chinese and Manchus were still tense. In many cases, he expressed his distrust of Chinese officials. Also, he found that Chinese scholars were incompetent in scientific matters. This is one of the key reasons why he studied European mathematics himself in order to win the admiration of Chinese officials. Since his youth, mathematics became an important part of his political life. As a ruler, Kangxi successfully enhanced Manchu prestige through science.

In fact, mathematics was a very useful tool for the emperor to show off his learning in his dialogues with his officials. Hence, out of fear that his newly acquired mathematical knowledge would spread beyond his control, he sometimes strove to keep it for a while as his "private property," as he did with the knowledge of algebra acquired from Antoine Thomas. This may well explain why some mathematics books translated into Chinese were not published during his lifetime, their knowledge therefore not spreading beyond the imperial court.

Theoretically, the Kangxi court should have provided Chinese with a good opportunity to learn more extensively from European sciences, since it had so many Jesuits who had close contacts with European scientists. However owing to the limits in his own understanding, his somewhat narrow perspective, and his desire to monopolize European knowledge, the Kangxi Emperor impeded the transmission of European mathematics and restricted it mainly to those within his imperial court. However, the publication of the *Basic Mathematical Principles* did benefit mathematicians generally in the late eighteenth and nineteenth centuries and led to the rediscovery and study of Song-Yuan mathematics. ([18]) In this way, the introduction of European mathematics during the Kangxi reign nonetheless played an important role in the history of Chinese mathematics outside the court as well.

Acknowledgements. The author is grateful to Prof. Karine Chemla for her invaluable comments and suggestions on an earlier draft of this article, to Prof. Noël Golvers for sharing me with his research on Antoine Thomas, and to Prof. Joseph McDermott, Prof. Joseph W.

Dauben and W. Kang Tchou for help in "sprucing up" my English. The author also gratefully acknowledges the support of K. C. Wong Education Foundation, Hong Kong.

References

[1] Aiton, E.J., *Leibniz: A Biography*. Bristol, 1985.

[2] Chemla, K. *The History of Mathematical Proof in Ancient Traditions*. Cambridge: Cambridge University Press, 2012.

[3] Golvers, N., *Libraries of Western Learning for China. Circulation of Western Books between Europe and China in the Jesuit Mission (ca.1650–ca.1750)*. vol.2, Formation of Jesuit Libraries, Leuven: Leuven University Press, 2013.

[4] Qi Han, 康熙时代传入的西方数学及其对中国数学的影响 (*The Introduction of Western Mathematics during Kangxi Period and its Influence on Chinese Mathematics*), Ph.D. thesis, Institute for the History of Natural Science, Chinese Academy of Sciences, Beijing, 1991 (in Chinese).

[5] _____, "君主和布衣之間：李光地在康熙時代的活動及其對科學的影響"(Between the Emperor and Mathematician: Li Guangdi's Activity during the Kangxi Reign and Its Influence on Science), 清華學報 (*Tsing Hua Journal of Chinese Studies*)(Hsinchu), New Series, 26:4 (1996), 421–445 (in Chinese).

[6] _____, "Patronage Scientifique et Carrière Politique: Li Guangdi entre Kangxi et Mei Wending", *Etudes Chinoises*, 16:2 (1997), 7–37.

[7] _____, "白晉的《易經》研究和康熙時代的西學中源說" (Joachim Bouvet's Study of the Yijing and the Theory of Chinese Origin of Western Learning during the Kangxi Era), 漢學研究 (*Chinese Studies*)(Taipei), 16: 1 (1998), 185–201 (in Chinese).

[8] _____, "格物穷理院与蒙养斋：17、18 世纪之中法科学交流"(L'Académie Royale des Sciences et les activités scientifiques en Chine aux XVIIe et XVIIIe siècles), 法国汉学 (*Sinologie Française*)(4),Beijing, Zhonghua Pulishing House, 1999, 302–324 (in Chinese).

[9] _____, "Emperor, Prince and Literati: Role of the Princes in the Organization of Scientific Activities in Early Qing Period", in Yung Sik Kim & Francesca Bray ed., *Current Perspectives in the History of Science in East Asia*, Seoul: Seoul National University, 1999, 209–216.

[10] _____, "The spirit of self-dependence and the appropriation of Western science: The transition in Chinese literati's attitudes toward Western science and its social context (ca. 1700-1760)" ('自立' 精神与历算活动 —康乾之际文人对西学态度之改变及其背景), 自然科学史研究 (*Studies in the History of Sciences*), 21:3 (2002), 210–221 (in Chinese).

[11] _____, "Antoine Thomas, SJ, and his Mathematical Activities in China: A Preliminary Research through Chinese Sources", in *The History of the Relations Between the Low Countries and China in the Qing Era* (1644-1911), ed. W. F. Vande Walle, Leuven: Leuven University Press, 2003, 105–114.

[12] _____, "L'enseignement des sciences mathématiques sous le règne de Kangxi (1662-1722) et son contexte social", in *Education et Instruction en Chine. II. Les formations spécialisées*, eds. Christine Nguyen Tri and Catherine Despeux, Paris-Louvain: Editions Peeters, 2003, 69–88.

[13] Qi Han and Jami, C., "康熙时代西方数学在宫廷的传播 —以安多和《算法纂要总纲》的编纂为例" (The Circulation of Western Mathematics at the Court during the Kangxi Period: A Case Study of the Compilation of the *Suanfa Zuanyao Zonggang* by Antoine Thomas), 自然科学史研究 (*Studies in the History of Sciences*), 22: 2 (2003), 145–155 (in Chinese).

[14] Qi Han, "A French Model for China: The Paris Academy of Sciences and the Foundation of the *Suanxue guan* (Academy of Mathematics)", paper presented to symposium "Science under Louis XIV and under Kangxi: a comparative approach to state policies and exchanges" (July 28, 2005, co-organized by Catherine JAMI and HAN Qi), XXII International Congress of History of Science, 24–30 July, 2005, Beijing. http://sourcedb.ihns.cas.cn/cn/ihnsexport/200906/W020140330376419936400.ppt

[15] _____, "1713: A Year of Significance", a lecture presented at REHSEIS, CNRS, Paris, 9 January 2007. http://sourcedb.ihns.cas.cn/cn/ihnsexport/200906/W020140330376419936400.ppt

[16] _____, "康熙时代的历算活动：基于档案资料的新研究" (Mathematical and Astronomical Activities during the Kangxi Reign (1662–1722)—A New Approach through Archival Documents), in Zhang Xianqing ed., 史料与视界：中文文献与中国基督教史研究, Shanghai: People's Publishing House, 2007, 40–60 (in Chinese).

[17] _____, "Between the Kangxi Emperor (r.1662–1722) and Leibniz: Joachim Bouvet's (1656-1730) Accommodation Policy and the Study of the *Yijing*", in Shinzo Kawamura & Cyril Veliath eds. *Beyond Borders: A Global Perspective of Jesuit Mission History*. Tokyo: Sophia University Press, 2009, 172–181.

[18] _____, "The transmission of Western mathematics and the revival of Chinese traditional mathematics in the Qianlong-Jiaqing period (1736–1820)"(西方数学的传入和乾嘉时期古算的复兴 —以借根方的传入和天元术研究的关系为例), in Chu Pingyi ed., 中国史新论：科技与中国社会, Taipei: 联经出版社, 2010, 459–486.

[19] _____, "科学、知识与权力 —日影观测与康熙在历法改革中的作用" (Science, Knowledge and Power: Observations of the Shadows of the Sun and the Kangxi Emperor's Role in the Calendrical Reform), 自然科学史研究 (*Studies in the History of Sciences*), 30:1 (2011), 1-18 (in Chinese).

[20] Jami, C., *J.-F. Foucquet et la modernisation de la science en Chine, la "Nouvelle Méthode d'Algèbre"*. Mémoire de maîtrise, Université de Paris VII, 1986.

[21] _____, "Learning mathematical sciences during the early and mid-Ch'ing", in B. Elman & A. Woodside eds., *Education and Society in Late Imperial China 1600–1900*. Berkeley: University of California Press, 1994, 223-256.

[22] Jami, C. and Qi Han, "The Reconstruction of Imperial Mathematics in China during the Kangxi Reign (1662–1722)," *Early Science and Medicine: A Journal for the Study of Science, Technology and Medicine in the Pre-modern Period*, 8: 2 (2003), 88–110.

[23] Landry-Deron, I., *Les leçons de sciences occidentales de l'empereur de Chine Kangxi (1662–1722): Textes des Journaux des Pères Bouvet et Gerbillon*, Mémoire du Diplôme de l' EHESS, Paris, 1995.

[24] Peng, Rita Hsiao-fu, "The K'ang-hsi Emperor's absorption in Western mathematics and astronomy and his extensive applications of scientific knowledge", *Li-shih Hsüeh-pao* 3 (1975), 349–422.

[25] Mme Yves de Thomaz de Bossierre, *Un Belge mandarin à la cour de Chine aux XVIIe et XVIIIe siècles*, *Antoine Thomas 1644–1709*. Paris, 1977.

[26] von Collani, C., *P. Joachim Bouvet S.J. Sein Leben und sein Werk*. Nettetal: Steyler Verlag, 1985.

[27] Widmaier, R. (ed.), *Leibniz korrespondiert mit China: der Briefwechsel mit den Jesuitenmissionaren (1689–1714)*, Frankfurt am Main: Vittorio Klostermann, 1990.

[28] Witek, J. W., *Controversial ideas in China and Europe: A Biography of J.-F. Foucquet (1665–1741)*. Rome, 1982.

Institute for the History of Natural Sciences, Chinese Academy of Sciences, 55 Zhongguancun donglu, Haidian District, Beijing 100190, China

E-mail: qihan63@hotmail.com

One hundred years after the Great War (1914–2014)

A century of breakdowns, resumptions and fundamental changes in international mathematical communication

Reinhard Siegmund-Schultze

Abstract. The paper describes and analyzes changing political, social and institutional conditions for international mathematical communication during the last one hundred years. The focus is on the Western Hemisphere and on relatively peaceful times between and after the two wars. Topics include the boycott against German and Austrian science, Rockefeller support for the internationalization of mathematics, the mass exodus of mathematicians from Europe in the 1930s, the resumption of mathematical contacts after WWII, the growing awareness of mathematics in the Soviet Union, and the emigration of Russian scholars to the West before and after the Fall of the Iron Curtain. Some emphasis is put on the barriers of language and culture between European, American and Russian mathematics and on the influence of Bourbaki during various periods. Several decisive events from the history of the ICM and the IMU are mentioned for their bearing on international communication.

Mathematics Subject Classification (2010). 01A60, 01A61, 01A80.

Keywords. International mathematical communication, World War I and II, emigration of mathematicians.

1. Introduction

When the German Nazis rose to power in 1933 the mass emigration from Europe, above all of the Jewish people, began. In the context of immigration the secretary of the American Mathematical Society, Roland Richardson said in June 1934:

> *Since the war, we have been constantly compelled to think of colleagues as nationals and not as citizens in the international domain.* [41, p. 16]

It is well-known that the rise of the Nazis can be related to that key catastrophe of the 20th century, World War I, which had poisoned international relations also more broadly.

We are gathering exactly one hundred years after the outbreak of that war. The mathematical world was smaller then, more focused on Europe and the United States. Our discussion will still have a bias towards the Western Hemisphere which is less justified for recent decades in view of the development of mathematics in Asia, South America and elsewhere.

We will try to give a short overview of major changes in the social and political conditions that affected international mathematical communication during the past one hundred

Proceedings of the International Congress of Mathematicians, Seoul, 2014

years. Global political events and strategies came of course quite often in ideological and philosophical wrapping and in this respect there are connections of 'internationalism' in scientific and mathematical communication with educational ideals, with notions of modernity in society, particularly due to the growing social impact of mathematics. An example of the connection between modernity ideals in mathematics and science and ideals in society and culture is the young group of French mathematicians in the 1930s, called Bourbaki, whose goal of an internationally aware, modern mathematics, was paralleled by programmatic attitudes in other social, cultural and political domains.[1] The use of natural languages, in particular the growing importance of English, is an eminently political phenomenon with economic context, and is at the same time intimately connected to communication in the sciences. The most recent striking changes in the conditions of international communication in mathematics are of course due to increasing globalization and the digital revolution. While much has changed, there are some invariants too, which are maybe even the more surprising, in particular a certain stability of national educational ideals.

One major conclusion which can be drawn from the past one hundred years is that international mathematical communication depends on political conditions and power constellations, and that mathematicians must for this very reason carve out – by active organizational and political engagement – such working conditions, which are least vulnerable to abrupt political changes. One consequence has been the re-definition in 1985 of the notion of a "country" in the statutes of the International Mathematical Union (IMU), which now say in §4:

> The term 'country' is to be understood as including diplomatic protectorates and any territory in which independent scientific activity in mathematics has been developed, and in general shall be construed as to secure the broadest and most effective participation of mathematicians in the scientific work of the Union.[2]

It therefore seems reasonable in the following discussion to use 'international' in the sense of 'inter-country' relations in the broader meaning of 'country' as outlined in the revised statutes of the IMU.

A complete history of mathematics in the 20th century which does justice to its enormous technical complexity has yet to be written. First approaches have been made both by historians of mathematics and working mathematicians.[3] Also in the following no detailed discussion can be given of 'mathematical communication' between individuals and schools in mathematical research,[4] in teaching and application, and we take for granted the changes

[1] Former 'Bourbakist' Pierre Cartier acknowledges the ideological dimension in Bourbaki, which was personified by some of its members at the time, such as Jean Delsarte and André Weil. See [36, pp. 26–27]. Cartier explicitly mentions the manifesto of the surrealists as a parallel event. See [22] for more details from that perspective. Liliane Beaulieu, in her unpublished dissertation of 1990, provides much on the social and ideological background of Bourbaki, information which cannot be found elsewhere in the literature.

[2] Statutes of IMU Approved by the General Assembly on August 16, 2010. Olli Lehto, the longstanding secretary of the IMU, describes in detail [20, pp. 245–250] why the IMU, which continues to have "international" in its name, was forced in 1985 to delete in its statutes the attribute "national" from its "adhering organizations" in order to enable both the People's Republic of China and Taiwan to participate.

[3] See [2, 6–9, 11, 16, 23, 26, 27].

[4] The standpoint of the outsider vis-à-vis national schools often triggered fruitful developments. Just to give two examples: The Austrian Ernst Fischer and the Hungarian Frigyes Riesz combined results of the French and German schools in the Riesz-Fischer Theorem (1907). In the late 1930s the Frenchman André Weil saw and utilized deeper connections between German arithmetical research and the theory of geometric correspondences of the Italians [34].

in the technical means of communication in the past hundred years: with respect to the latter a reference to air traffic and email should suffice. However, 'international mathematical communication' is much more than the two aspects just mentioned and it includes, for instance, international comparisons which inform national developments and induce change in 'national mentalities'. We will go into some of these conditions and relevant political events, and provide literature for further and detailed study. Applied mathematics performed in non-academic environments, hybrid disciplines such as aerodynamics and statistics, which are not core fields at the present ICM, are not covered in the following discussion. Let us just say that in these fields and their institutions usually exist very specific problems with respect to international communication.[5]

Let me first make some tentative and preliminary remarks about the current situation in international communication.

The identity of the mathematician has greatly changed and the social role of mathematics has strongly increased worldwide since the Great War, the latter for instance evidenced by the digital revolution. However, primary and secondary school teaching in mathematics and undergraduate university education remain predominantly national (country) tasks. Also research is mostly supported by national institutions, in spite of the existence of supranational agencies such as EU, OECD, and UNESCO, and of internationally acting private enterprise, and even though the individuals thus supported are not confined in their origin to the given national system. The global political situation, after Nine-Eleven, in the Near East, Ukraine, and some parts of Asia is by no means safer than it was at the outbreak of the Great War a hundred years ago. Science, including mathematics, has to accept its part of responsibility for the development of new weapons and other technical means of conflict and destruction. After the fall of the Iron Curtain (and partly caused by it), developing countries like China, India and Brazil aspire for mathematical prowess in a similar way to newborn nations like Poland, and economically and politically growing ones such as the United States and Soviet Russia after World War I. There is no dearth of new and old nationalisms, including appeals for scientific boycotts (conjuring up the memory of the boycott after the Great War). Particularly in the past 25 years, post-modernist indifference has given way to renewed fundamentalism and anti-scientific resentment, while justified and nuanced criticism of science and its impact on society has increased as well.

International comparisons of educational and science systems continue to be on the agenda, exemplified by PISA and TIMSS, even though some of these comparisons are disputed in their methodology and predictive value. The United States remains the only scientific superpower, somewhat similar to the predominance of Germany around 1914. This is also demonstrated by the figures for invited talks at the present congress: 64, i.e., about one third of the 190 sectional speakers come from U.S. institutions, while roughly another sixth and twelfth come from French and UK-institutions respectively, adding up to over a half of the total of invited speakers. That only four of the sectional and none of the plenary speakers is currently employed by Russian institutions does not mirror the representation of Russian-born (and Russian-socialized) mathematicians at the present congress, which is apparently much stronger.

[5] The reference to big companies such Microsoft, Google and Apple, which act internationally but are still based on national systems with all legal repercussions and implications for corporative and national loyalties (case E. Snowden) might suffice to warn against an extension of our discussion into this direction. Keyfitz, in an article on industrial mathematics published in connection with the Madrid ICM in 2006 [18], emphasizes the differences in the identity of the industrial mathematician compared to the academic mathematician.

A somewhat similar picture is presented by the dominance of American institutions in relation to the origin of many Fields medalists of recent decades, again with French institutions ranking second.[6]

At least two things seem to have greatly changed in the 21st century.

First, in a globalized world attacks on and defenses of mathematics are not usually based on nationally sanctioned ideological tenets, as was the case in the 1930s when slogans such as "mathematics is a function of race" or "set theoretic mathematics is bourgeois idealism" figured in political campaigns.[7]

Second, there seems to be a leveling of what once were considered to be recognizable national styles within mathematics [39], while there continue to exist recognizable styles of internationally structured research schools: this is of course partly connected to the leveling of natural languages in mathematics, the adoption of English as the lingua franca in mathematics and other fields of science and continuing emigration of students and trained mathematicians. Much could be said about communication under war conditions – or partly the absence of it – and the influence of military funding and secrecy regulations on international mathematical collaboration. Research on this question is gradually developing.[8] Of course, in times of war national, as opposed to international, communication dominates. However, international comparisons, mentioned above, are crucial at these times too. The military funding of mathematics and resulting implications for international collaboration also in relatively peaceful times has repeatedly caused discussions among mathematicians in various countries. For the U.S. we have, for instance, the following statement by a leading mathematical educationalist:

> *Many of us are military-oriented because of the long involvement of mathematics in military science. But we also tend to be internationalists since mathematics is an international culture, independent of language and politics. This characteristic has sometimes got us into trouble, or at least made us suspect as security risks.* [14, p. 430]

With a focus on times of relative peace and 'normality' of international communication it seems reasonable to divide the past hundred years roughly in three main periods: Interbellum,[9] Postwar and Cold War, and, thirdly, the ongoing Information Technology Revolution, the latter partly triggered by (in the former East) and partly triggering the fall of the Iron Curtain. My discussion concentrates on the first period of the three since this period has been subject to the greatest amount of historical research.

2. Interbellum

The Great War had deep consequences for the international landscape in mathematics and for the emotions and mentalities of mathematicians. While the concrete impact of nationalistic

[6] When we talk about American, we are referring to the United States.

[7] The 20th century, from 1917 to 1989, has been, in the opinion of many, the ideological age. See also Cartier in [36, p. 27].

[8] See [6], and in it [44], as an approach to an international comparative perspective and providing more literature. See also [15].

[9] Given the mass exodus triggered by the Nazi rule in Germany one could argue that this period be divided in two: before and after 1933.

sentiment stirred by the war on world-wide research in mathematics is difficult to evaluate, the more material consequences of the Great War are not. On the French side, a large number of the students of the École Normale Supérieure had perished in the war [3]. New nations appeared on the political scene: Polish mathematics became a power house of set theory and functional analysis during the 1920s and 1930s. The experiences of the War had shown the growing importance of science in the competition of nations. This had repercussions for mathematical institutions, for instance the foundation of journals, series of monographs (Springer)[32], and institutes for applied mathematics.[10]

In the U.S. and Russia, with enough resources of their own, the War (accompanied on the Russian side by a political revolution and its consequences) led, by different roads, to a shared world dominance in mathematics (which was finally accomplished after World War II).

In the United States, plans designed immediately after the War for developing the scientific publication system ([38, 53]) and applied mathematics ([43]) were temporarily postponed in favor of developing personnel and strengthening research. The War had put an end to the tradition of Americans going to Europe, in particular to Germany and France, for study [31]. The impoverishment of European states, not just the defeated ones, and political unrest in the Soviet Union created unique possibilities for recruiting first class personnel from Europe.[11] The superior material strength of the U.S. (above all the system of private universities) led to a first and early wave of emigration, which – combined with the effects of the forced emigration from Europe after 1933 – bore fruit some twenty years later. But American postdocs went also in increasing numbers in the opposite direction, to Europe, mainly supported by the Rockefeller and Guggenheim philanthropies. American mathematics in general became much more research conscious than before the war. At the same time Asian students (partly supported by the Boxer Indemnity Scholarship Program predating WWI) began to appear in greater numbers in the U.S., and soon exceeded the number of European undergraduate students there [53, p. 492].

Although the Soviet Union faced growing political isolation, and a loss of scientific personnel and although the language divide persisted, contacts with French [13] and German mathematicians flourished there before 1933, for instance in Göttingen.[12] We will see below, that the relationship of the Russians to the Americans, in particular to the Rockefeller Philanthropy, remained strained. But due to its – compared to the U.S.– broader and older mathematical traditions and due to a critical scientific mass and size of its own, Soviet mathematics was able to develop strongly even in the 1930s and 1940s when its international isolation once again increased.

In Western Europe the 'boycott of German and Austrian science' organized by the various international scientific unions (including the mathematical one), which had been founded in the aftermath of the war, had a great impact on the emotions of scientists and on the official channels of communication; as is well known Germany and Austria were excluded from the international congresses of mathematicians in Strasbourg (1920) and Toronto (1924). However, it is less clear what impact this actually had on the informal mathematical collaboration

[10] In Poland the first specialized international journal of mathematics appeared: *Fundamenta Mathematicae* (1920). Institutes for applied mathematics were founded in St. Petersburg by V. Steklov (c.1920), in Berlin by R. von Mises (1920), following older traditions in Göttingen under C. Runge (1904), and in Italy by M. Picone (1927).

[11] Mathematicians who emigrated from Russia such as S. Lefschetz, N. Minorsky, J. Shohat, J. Tamarkin, S. Timoshenko, often had interests and training in applications.

[12] P.S. Aleksandrov was a frequent guest.

between individuals and mathematical schools, although singular drawbacks cannot be denied.[13] Many informal networks of collaboration survived the war (such as the one between G.H. Hardy in England, Harald Bohr in Denmark and Edmund Landau in Germany) or were even extended (with Copenhagen becoming a hub for international collaboration in mathematical physics), and some Germans felt that those among the mostly French and Belgian mathematicians (such as Émile Picard) who insisted on the boycott, were increasingly isolating themselves. Arguably, by the late 1920s and early 1930s, mathematics in Germany, particularly in Göttingen, had become the most 'internationalized' of all national mathematical cultures in the world. This was true relative to a number of metrics: the nationality and origin of mathematicians teaching and studying at German universities; the number of German mathematicians sent abroad either as postgraduate students via, for example, Rockefeller's International Education Board, or as guest professors like Richard Courant, and Wilhelm Blaschke; national origins of authors of articles in German journals; the international importance of the German publication system in mathematics; and the variety of topics discussed.

And still, Germany was no longer as dominant a mathematical world power as it had been before the war. There were shortcomings in algebraic topology which, as a 'modern' and axiomatic mathematical sub-discipline, had reached firm ground in the U.S., particularly at Princeton, since the early 1920s. Similar remarks concern lack of work in functional analysis and real functions; this was partly reinforced by reservation on the German side against contacts with Polish and French mathematicians, resentments which did not exist on the part of the Austrians (H.Hahn, E. Helly, E. Fischer etc.). In probability theory, the French and Russian schools were much stronger than the German one, Richard von Mises being essentially the only contributor on the German side during the 1920s. The English and Scandinavian, and somewhat later American, schools took the lead in mathematical statistics.

A look at international mathematical communication should not be restricted to research systems but has to include the development of school systems and educational ideals, which played an increasing role within developing mass education, affecting norms and standards in mathematics as well; international comparison was important in these developments because it inspired national developments, even though often merely in a propagandistic manner by exaggerating or distorting foreign investment and accomplishments.[14] The developing U.S. school system has often been criticized by American educators themselves, both before and after World War I, for unwillingness to learn from foreign (European) experience. W.L. Duren, himself very much involved in educational policies, found in 1989, that the U.S. in 1918 "set forth the agenda of social development and personal fulfillment as the aims of secondary education, and relegated the mastery of subjects to low priority" which in his opinion resulted in a growing "isolation from European Mathematics" [14, p. 405].

It would take a broader discussion of the history of mathematical education (which is not intended here), whether mutual "isolation" in Duren's sense, meaning self-contained development of educational systems, was not partially an international phenomenon, and not restricted to the U.S. Suffice it to say that the "mastery of subjects" by students and

[13] Bru observes [7, p. 176] that the ICM in Bologna 1928 was a missed opportunity to establish connections between parallel French (J. Hadamard, M. Fréchet), Czech (B. Hostinský) and German/Austrian (R. von Mises) work on Markov chains. Out of political resentment von Mises had chosen not to go to Bologna and remained unaware of the presentations given there for several years. Conversely, the French did not learn about the progress which von Mises had made with the help of the theory of positive matrices (G. Frobenius).

[14] Discussing the 19th century, Parshall finds that "educational reform ... represented a sort of international common denominator in the formation of these national mathematical constituencies." [25, p. 1581].

teachers themselves remains hotly debated today on an international scale.

The controversial discussion of educational ideals leads back to the influence of the Rockefeller Philanthropy, whose main international agency in the 1920s was the "International Education Board" [41]. Its twofold aim was helping European mathematics to recover and American mathematics to develop and broaden, however, rather in the elitist sense of "making the peaks higher" than relegating the "mastery of subjects to low priority." Rockefeller's support for mathematics is probably world-wide the most important factor for internationalization in the 1920s, only comparable with the effect of the mass exodus from Europe after 1933. It was affected through Rockefeller's international fellowship program and Rockefeller's financing of the new mathematical institute buildings in Paris (later to be called Institut Poincaré) and Göttingen, which opened respectively in 1928 and 1929. Although the Rockefeller people, advised both by Americans (O. Veblen and G.D. Birkhoff) and Europeans (R. Courant, É. Borel, H. Bohr etc.) did not have short- sighted goals such as the brain-drain of European mathematicians, the priority of national American values and developments in the eyes of the philanthropy was never in doubt. The predominance of private money in funding American mathematics until WWII was epitomized by the name "National Research Council Fellowships" for the Rockefeller-financed grants reserved for American candidates, who usually had to pass lower quality standards than European candidates. Rockefeller policies revealed a clear preference for American and Western European fellows over Eastern European, in particular Russian ones, who were only occasionally and indirectly (through West European sponsors) supported. The philanthropy did not originally reach out beyond Europe and North America either. As late as 1932, for instance, a Rockefeller officer said that "we are not permitted to consider subjects from India." [41, p. 222]. It was only in the mid-thirties and finally during WWII that – for obvious political reasons – South American candidates were increasingly supported by Rockefeller and other American philanthropies.

The focus of Rockefeller support on Göttingen sharpened jealous institutional conflicts within Germany, in particular with mathematicians in Berlin (L. Bieberbach, E. Schmidt, R. von Mises) who were skeptical of some traits of modernization, in particular internationalization. These included developments in commercial publishing, when for instance some mathematicians half-jokingly called the Springer Grundlehren series the 'yellow peril'. This overlapped with concerns for content and language, for instance in mathematical reviewing, where Springer's new *Zentralblatt*, founded in 1931, published abstracts in English unlike its older rival *Jahrbuch über die Fortschritte der Mathematik* [38].

At the same time developments within mathematics, influenced by David Hilbert's axiomatic method, supported internationalization of mathematics, based on commonalities in the mathematics of various national schools. The American E.T. Bell called "abstract spaces" a "typical example of the internationalism of mathematics" [2, p. 543]. Again it is a difficult problem to decide whether Emmy Noether's school of abstract algebra in Göttingen in the late 1920s primarily fostered internationalization or whether it was already a result of it. Certainly, a particular abstract and structural style of presentation, as in the famous book of Noether's student B. L. van der Waerden, *Moderne Algebra*, made it easier for the discipline to cross boundaries of language and mentality, to 'internationalize.' Indeed, between 1932 and 1935 several American (Garrett Birkhoff, Saunders Mac Lane) and French (Jean Dieudonné, Henri Cartan) mathematicians witnessed – by their own testimony – a kind of quasi-religious 'conversion' toward abstract algebra in the Noetherian sense. This marked the beginnings of the French group of young mathematicians, Bourbaki, which later became

the quintessential propagandist of the structural approach in various mathematical disciplines [11].

For all the internationality reached in Göttingen's mathematics towards the end of the 1920s there remained a feeling even among their leading figures that not everything was secure and irreversible in international mathematical communication. Throughout the 1920s nationalist resentments persisted among various European scholars and students, above all in Germany, towards a revision of the results of World War I. In Göttingen anti-Semitic actions, particularly in the student body, increased political tensions. In the mind of Hilbert and other modernist and internationalist mathematicians this political insecurity contributed to maybe exaggerated concerns about a possible anti-Cantorian backlash in the logical foundations of mathematics [22]. This was the case not least because the principal opponent of mathematical formalism, the Dutch intuitionist amd topologist L.E.J. Brouwer, seemed to personify both the mathematical and the political counter-revolution. Together with nationalistic German mathematicians such as Ludwig Bieberbach, Brouwer opposed the participation of German mathematicians in the ICM of Bologna in 1928 out of concerns for German national pride. Although Bologna finally saw the reappearance of a German delegation, the intended and actual[15] presentations of old and frail Hilbert at the congress expressed his double concerns and contain some element of desperation, maybe aggravated by the state of his health.[16] In a 3-page political talk Hilbert intended to say at Bologna:

> *It is a complete misunderstanding of our science to construct differences or even incompatibilities according to peoples and races, and the reasons for which this has been done are very shabby ones. Mathematics knows no races. ... For mathematics, the whole cultural world is a single country.*[17]

At Bologna mathematicians agreed that there were still international conflicts and therefore decided to reconvene in the politically neutral surroundings of Zürich in 1932.[18]

The following year, 1929, saw the beginning of the Great Depression, and even the resourceful Rockefeller Philanthropy had to reduce its activities in Europe, focusing from now on even more on the U.S. The seizure of power by the German Nazis in 1933 brought international mathematical communication for mathematicians within Germany almost to a standstill. Bieberbach, who had become a Nazi, would soon speak deprecatingly about "international formalism" [38, p. 320] in mathematics.

Under their political regimes in the 1930s, both German [35] and Russian mathematicians became internationally isolated towards each other and towards the West. They had

---

[15] Hilbert presented a plenary talk in Bologna on "Problems of the foundation of mathematics" where he reiterated his famous "In mathematics there is no ignorabimus", already known from his talk at the ICM in Paris 1900.

[16] Even the state of his health had political connotations, because Hilbert was finally cured by American medication (provided through the Rockefeller philanthropy) while the renowned German health system had been unable to help him.

[17] Although this passage has been quoted all over the place as part of an actual talk given by Hilbert (e.g. in [20, p. 48] and even though it follows closely a manuscript in the handwriting of Hilbert's wife Käthe, I have so far no reason to believe that this additional, political talk was actually given. Neither the Proceedings of the Congress in Bologna nor any of the many published reports on the Congress mention Hilbert's talk, not even in connection with some social event of the congress. A detailed letter from Hasso Härlen to Brouwer, dated Eislingen, 27 September 1928, about Bologna does not mention it either, which is probably the most convincing counter-evidence so far, although a page or so is missing at the end of the letter. Thanks go to Dirk van Dalen (Utrecht) for providing a transcription of this letter.

[18] Härlen to Brouwer, 27 September 1928. This was, in a way, a duplication of the political decision to have the very first ICM in Zürich 1897.

their participation in foreign congresses curtailed by their regimes, their private mail was controlled, and their publication in international journals diminished. The influential international conference on topology in Moscow in September 1935 was about the last international event with participation of Russian mathematicians before the war. From 1934, Springer's abstracting journal *Zentralblatt* was edited from Copenhagen in order to avoid the possibility of political influence by the German government. The editor, who had had the post since 1931, was the German-Austrian refugee and historian of mathematics Otto Neugebauer. Neugebauer was also responsible for Springer's *Ergebnisse* ('Results') series, and in that context he wrote on 14 March 1937 to his friend Richard Courant, who was by then in New York:

> *You will certainly be interested to learn that Kolmogoroff and Khintchine had big scandals in Russia due to their Ergebnisse-reports, published in Germany. As a matter of fact, in Russia there is now flourishing the same idiotic nationalism as in the Third Reich. Of course you should not write about these things to Russia, but you ought to know because of the Yellow Books. For instance I do not believe that either one of the two would now be able to write a Yellow Book without danger.*[19]

Inside Germany and Russia the communication and the publications systems continued to work relatively undisturbed – in Germany of course only after the disruptive and shameful events of expelling Jewish mathematicians from their posts. The Russian publication system was highly subsidized by the state, and foreign literature was often published in pirated translations. The German mathematical publication system, headed by Springer, especially monographs and *Zentralblatt*, remained internationally strong throughout the 1930s. For the Hitler regime, which pursued policies of economic autarchy, it became a source of much coveted foreign currency.

Just as before, at the time of the boycott, 'internationalism' meant cooperation primarily between mathematicians from politically allied countries. The insurmountable dogma of anti-Semitism in German politics created additional problems for the international communication of German mathematicians. A striking example is the journal *Compositio Mathematica*, which had been founded around 1930 by Brouwer in the Netherlands, at that time supported by Bieberbach, Brouwer's ally in anti-boycott policies. The journal was expressly intended to further the development of mathematics and, at the same time, international cooperation. When *Compositio Mathematica* finally appeared for the first time in 1934, its international editorial board included several Jews, who had fled from Germany. This led to a withdrawal of German mathematicians from the board. The journal was suspended during the war when the Netherlands were occupied by German forces [21, p. 235].

The divided internationalism on the German side led to an expansion of contacts of German mathematicians to Asia, renewing older contacts made in the 1920s, when scholars both from Germany (Th. v. Kármán, W. Blaschke) and from Western countries (J. Hadamard, N. Wiener) had assumed guest lectureships in the East. Of course there existed older contacts between Japanese and German mathematicians from around 1900 particularly in number theory through Teiji Takagi.[20] After 1933 Chinese students increasingly came to study in

[19] Courant Papers New York University Archives, no call number. My translation from German. Neugebauer was alluding to the two influential books on probability theory of Kolmogorov and A. Khinchin which appeared in 1933 and 1934 in German in the *Ergebnisse* series.

[20] Takagi, who had been studying in Göttingen in around 1900, wrote his most important paper in 1920 which

Germany. While the full story of German-Asian mathematical relations during the Nazi era is not yet documented historically, the importance of Chinese geometer S. S. Chern's stay in Hamburg from 1934 and his communication with Erich Kähler and Wilhelm Blaschke have been repeatedly stressed.[21]

Particularly in the case of Italy, political alliances made collaboration for German mathematicians easier [42]. The leading researcher in algebraic geometry and politically well-connected Italian mathematician Francesco Severi said at the end of his talk at a conference in 1938 in German Baden-Baden:

> *I hope that the important progress that Germany has realized in modern algebra, will allow her magnificent mathematicians to penetrate deeper and deeper into algebraic geometry which has been cultivated in Italy over the past 40 years; and that the connections between German science and Italian science, which have already been so close in this domain at the time of our masters, become more intimate every day, as they are today in the political and general cultural realm.* [34, p. 15]

Severi's main partner on the German side, H. Hasse, declared in 1939 vis-à-vis American mathematicians that there was a "state of war between Germans and Jews," thus supporting the introduction of policies into the *Zentralblatt* which prevented German-authored papers being reviewed by Jewish mathematicians. Somewhat later, during the war, Hasse tried to involve French mathematicians under occupation and some French prisoners of war in collaboration with the Germans [42].

The increasing division of internationality showed clearly at the Oslo ICM of July 1936. Italian mathematicians were forbidden by the Fascist regime to participate. This was a reaction to the international boycott of Italy, following the Italo-Abyssinian war and the occupation of Ethiopia by Italy in May 1936. Russian mathematicians were also prevented from coming; their participation having been a specific point on the agenda of a meeting of the Politburo of the Communist Party.

Mathematics at the time of Oslo still bore all the marks of 'little science' in the words of D.J. de Solla Price [28]. Compared with today, there was a relatively small attendance of a few hundred mathematicians at the International Congresses and a limited number of countries participating in them. The numbers had actually gone down, apparently due to the economic situation around 1932 and the political situation around 1936, from 836 in Bologna, through 667 in Zürich to officially 487 in Oslo.[22] At the same time, the 1930s saw an increase of smaller, specialized international mathematical meetings (topology, foundations, probability, applied mathematics), a fact which had, of course, implications for the decisions of mathematicians, in particular their willingness and ability to attend big congresses in addition to small ones.[23] After the war, international conferences on specialized

introduced the Takagi class-field theory generalizing Hilbert's class field. Hasse included Takagi's theory in his treatise on class field theory a few years later. Also a lectureship in Japan (1923–1928) of the German mathematician Wilhelm Süss should be mentioned because Süss became influential in German mathematics in the 1930s.

[21] Tobies lists 14 mathematics PhD students in Germany from China for the period 1907-1945, 13 of whom received their degree after the Nazis had come to power in 1933, several during the war [47, p. 18]. However, somewhat surprisingly, none from Japan is listed.

[22] Even the 487 participants recorded in the Oslo Proceedings was an exaggeration, because the Russian and Italian delegates listed did not attend. It was 1950 in Cambridge, MA, that for the first time over 1000 mathematicians took part, namely about 1700 [12, p. 151].

[23] Details about this development are given in [51, p. 316].

topics, developed as one of the main activities of the new IMU.[24]

On the other side of the political divide, in the traditional Western countries, the mass exodus from Germany, and later from German occupied territories, particularly Austria and Poland, brought about a total reshuffling of international communication, a strong increase in oral communication, changes in research subjects, in teaching, and in mentalities. While the emigrants had to adapt to the new environments, mathematics in the host countries, above all in the United States and in Great Britain, gained considerably [45].

For mathematicians who hitherto had been for the most time divided by the Atlantic, it was refreshing to experience new oral communication. Abraham Flexner, the director of the Princeton Institute for Advanced Study, wrote in April 1935 to a Rockefeller official about one of the internationally leading mathematicians of the time, the German Carl Ludwig Siegel:

> *Siegel . . . made a very deep impression upon the mathematicians here. They obviously knew of him while he was still in Frankfurt, but I don't think that they realized how able he was until they had the opportunity for closer personal contact.* [41, p. 197]

This reminds us of a quote from André Weil's historical plenary talk at the 1978 ICM at Helsinki: "We all know by experience how much is to be gained through personal acquaintance when we wish to study contemporary work; our meetings and congresses have hardly any other purpose." [52, p. 229]

As to the confrontation of different research mentalities during immigration, George Birkhoff's talk at the semi-centennial celebration of the AMS in 1938 is revealing. He said among other things that American research on what he called "Special Analysis"[25] had not been very widespread, because Americans tended "to take our mathematics as serious business (while) . . . many of the most astonishing mathematical developments began as a pure jeu d'esprit."[5, p. 307]. The Polish immigrant Stanislas Ulam, who was then at Harvard, considered this as a sign of "lack of self-confidence" and said it "was strange to me". However, he continued: "it was less objectionable than the European arrogance" [48, pp. 87–88]. Together with "Special Analysis", concrete classical analysis in a broader sense was introduced to the U.S. by some immigrants. Two American students of Polish analyst and immigrant Antoni Zygmund wrote in 1989:

> *He [Zygmund] realized that fundamental questions of calculus and analysis were still not well understood. In a sense, he was 'bucking the modern trends'.* [10, p. 347]

Finally, under war conditions, long experiences in cooperating with state bureaucracies, with the military, and industrial environments made immigrants such as Courant and Theodore von Kármán inspiring partners for American mathematicians, who traditionally had mostly functioned in purely academic environments [43].

Amongst all these gains and mutual profits, the losses of the exodus from Europe should not be forgotten. These occurred foremost at an individual level. Mathematicians had been

[24] See [20, p. 170]. The ICM at Cambridge, MA in 1950 had – in addition to the usual program – specialized conferences on algebra, applied mathematics, analysis, and topology. Chandler and Magnus discuss the importance of international specialized meetings in a special chapter in their book on *The History of Combinatorial Group Theory*, devoted to "modes of communication" [9].

[25] Birkhoff mentioned N. Wiener on Tauberian theorems, E. Hille, J. Tamarkin and D. V. Widder on Laplace–transformations and L. L. Silverman on summation of divergent series.

rooted up from their scientific and personal environments. Others did not make it and were killed back in Europe, for instance in Nazi camps. Until this day unpublished manuscripts are found in papers left by victims of the purge. Especially during the war there existed restrictions and secrecy regulations for immigrants from enemy countries, even if they had been expelled from those countries. The internment of German immigrants in Britain and of Japanese in the U.S. is well known. Mathematical communication even between allies, such, as the British and the Americans, was temporarily disturbed.[26]

Also the losses for mathematics in the deserted mathematical environments in Europe should not be ignored. These losses partly resulted from interrupted communication channels between Europe and the U.S. One could mention here the young and brilliant mathematician and fervent Nazi, Oswald Teichmüller, who was killed in the war, and whose works were temporarily forgotten.[27]

Much later, in 1977, the son of George Birkhoff, Garrett, an influential mathematician in his own right, spoke about the decisive new level of internationalization within the American mathematical community resulting from the developments of the 1930s and from the war. But he did not forget to add that at least some Americans (and he apparently included his father) viewed the impoverishment of the European scientific cultures around 1940 with mixed feelings and as potentially dangerous for the harmonic development of world science as a whole [4, p. 77].

3. Postwar-Cold War

Considerably less historical research has hitherto been done on international mathematical communication after World War II, than on the period before the war . Therefore the following remarks are by necessity less complete than those in the preceding section.

As a consequence of mass immigration and due to much increased state funding during and after World War II (much funneled through the department of defense and, since its foundation in 1950, through the National Science Foundation, NSF), the United States came out of the war as one of two mathematical super-powers, with the Soviet Union being the other. In his 1946 obituary of George Birkhoff, topologist Oswald Veblen from Princeton alluded to the AMS semi-centenary of 1938:

> *Among the unconscious revelations of the address on "Fifty years of American Mathematics," one of the most vivid is that of the depth and sincerity of Birkhoff's devotion to the cause of mathematics, and particularly "American mathematics." . . . It may be added that a sort of religious devotion to American mathematics as a "cause" was characteristic of a good many of his predecessors and contemporaries. It has undoubtedly helped the growth of science during this period. By now, mathematics is perhaps strong enough in the United States to be less nationalistic.* [49, p. xx]

One might add that both mass immigration and the experience of the Nazi crimes, particularly Auschwitz, had essentially eradicated xenophobia and - above all - anti-Semitism in

[26] See [44], [45].

[27] The so called 'Teichmüller-theory', disclosing deep connections between Riemann surfaces and quasi-conformal mappings, began to reappear after a publication of Lars Ahlfors in 1953. Its temporary neglect is partly due to the fact that important papers of Teichmüller's were published in the Nazi-journal *Deutsche Mathematik*.

American academia, sentiments which had still been very palpable in Birkhoff's talk before the AMS in 1938.

Former immigrants continued to help Americans in various ways to develop a new American mathematical culture. In the last months of the war immigrants to the U.S., such as Hermann Weyl and Richard Courant, and American mathematicians such as Arnold Dresden and AMS secretary John Robert Kline, discussed how to improve mathematical education in the U.S. and utilize the experiences of the immigrants in this process. In a letter to Weyl, Dresden emphasized that "we should discuss not merely the problems of graduate education but the entire range of mathematical education beginning with the elementary schools." On 18 February 1945, the four mathematicians mentioned and a few other Americans and immigrants[28] met at Dresden's institution, Swarthmore College, PA, and produced a memorandum where they criticized the level of teachers and their preparation, stressed the importance of the history of mathematics for education,[29] and emphasized:

> *It would also be useful to become acquainted with measures taken in foreign countries, particular England and Russia, for the betterment of mathematical education.*[30]

The group proposed the appointment of commissions "to study the matters presented in this report;" however, no immediate consequences of the report are known to this author.

The recovery of international contacts after the war was difficult for various resource-related and political reasons, including political mistrust (McCarthyism, Stalinism). Scientists had to find ways out of the secrecy regulations of war research, which soon were complemented by new ones at the beginning of the Cold War.

Again, early and recent immigrants to the U.S. were instrumental in this process. In 1963 the English analyst Mary Cartwright, who worked on non-linear vibrations and oscillations from the early 1940s, reported about restrictions of international communication even with the American allies during the war. She alluded to the language barrier between English and Russian, but also to the role of early immigrants to the U.S., such as Nicolas Minorsky, who helped to overcome that barrier:

> *While Littlewood and I were attacking special problems, Lefschetz, Levinson, Minorsky and others in the United States, impelled to a large extent by applications connected with the war, were beginning to prepare the way for a clearer unified and more easily handled mathematical theory. . . . Minorsky's book, 'Introduction to non-linear mechanics' was first published as a 'Restricted' report under the auspices of the U.S. Navy and appeared in parts between 1944 and 1946. This made the Soviet work more easily available to those who could obtain it, but the material was still very indigestible.* [8, pp. 196–197]

The resumption of international mathematical congresses and the re-foundation of the International Mathematical Union in 1951 have been described by Lehto [20]. He has stressed the role which Americans such as Marshall Stone and immigrants to the U.S. such as Courant played in this process. The contribution of victims of National Socialism who had remained

[28] H.W.Brinkmann, E. J. Miles, O. Ore, and H. Rademacher.

[29] Contrary to these proposals, history unfortunately finds only a marginal place in modern mathematical didactics.

[30] Weyl Papers, ETH Zürich, Hs 91: 196.

in Germany such as E. Kamke was also substantial.[31] Those individuals and the dismal experiences with the old Union ensured that a boycott of German mathematics after WWII was out of the question, in spite of all the justified bitter feelings in many countries about the Nazi crimes.

Several of the immigrants and also some other British, French and American mathematicians visited Germany [45]. This happened originally on post-war missions in order to evaluate German personnel and research during the war. In Germany the Mathematical Institute in Oberwolfach (Black Forest), which had been founded under the Nazis in 1945, became a place for the resumption of international contacts in mathematics. It is today an internationally well established and coveted place of international mathematical meetings, together with others, for instance Luminy (France) or the Banff International Research Station (Canada).

Nevertheless, at least in the beginning of renewed contacts there was – not unexpectedly in view of the open wounds of the war – plenty of misunderstanding between German mathematicians and foreigners, including emigrants. Restrictions were imposed on German research in more applied domains as a result of decisions by the Allied Control Council in Germany. One problem concerned the publication rights for German books seized by the U.S., where companies such as Dover republished large amounts of books during and after the War in original German, without paying royalties to the authors. However, it has been argued that the seizure of German books contributed to keeping German mathematics and the German language alive in the minds of the international community, at least for a while [40].

A second wave of emigration from Europe started after the war, not least caused by the precarious working and living conditions, especially in postwar Germany. The start of the brain drain from Europe to the United States not only affected German mathematics. Courant's Institute at New York University became a center of attraction for immigrants from several European countries.

The return to their home-countries of French and of some (if only a few) German emigrants after the war, and the rapidly increasing number of foreign students in the United States led to the importation and re-importation of certain mathematical sub-disciplines to Germany and to Europe. The influence of Bourbaki, has to be mentioned here. Many ideals of internationality and modernity in mathematics at least in the Western Hemisphere (for instance 'math = set theory') were partly mediated by Bourbaki in the decades to come; some influences on international educational ideals (New Math in various national forms) have been strongly criticized. There are even indications that the abstract, structural approach of Bourbaki deepened the divide between West European and Russian mathematicians.[32] This happened although in the 1930s Hilbert's axiomatic method had influenced both Bourbaki and Russian mathematicians, such as Andrey Kolmogorov.[33] In an interview of 1990, V.I. Arnold, a noted student of Kolmogorov, deplored the increasing distance of some parts of abstract mathematics from applications and went as far as saying:

In the last thirty years the prestige of mathematics has declined in all countries.

[31] Evidence for this can be found in the files of the American Mathematical Society at Brown University, Providence.

[32] One Romanian mathematician claims that adherence to Bourbaki was understood in some quarters in East Europe as a token of resistance again Soviet dominance, with old relations to France being part of the picture [33, p. 564].

[33] Think of Kolmogorov's axiomatics of probability of 1933.

> *I think that mathematicians are partially to be blamed as well (foremost, Hilbert and Bourbaki).* [19, p. 379]

Already at Harvard in 1950, at the postponed ICM which originally had been planned for 1940, German and Japanese mathematicians from the former enemy countries took part in roughly the same numbers (about 10 in each case) as other non-American countries. However, there were no participants from the Soviet Union or any other Socialist countries.[34] The East European countries joined the International Mathematical Union in the late 1950s, East Germany as an independent country only in 1964. The People's Republic of China did not become a member of the Union until 1986.

However, participation in the ICM did not necessarily require membership of the IMU, as Bologna 1928 had shown. The Russians returned to the international scene at the Amsterdam ICM in 1954. The importance of Kolmogorov's plenary, which enabled Western mathematicians such as Jürgen Moser to connect to largely unknown Russian research, has been repeatedly stressed [8]. Kolmogorov's lecture was entitled "General theory of dynamical systems and classical mechanics," and it was presented and published (in the Proceedings) in original Russian. While Soviet mathematician had given plenary talks in West European languages (mostly French and German) at earlier ICMs (for instance Nikolay Luzin in Bologna 1928), from the 1930s they had begun publishing almost exclusively in Russian, a practice which they continued after the war. This prompted the American Mathematical Society, with funding from the Office of Naval Research, to begin a Russian translation project in 1947. The Society of Industrial and Applied Mathematics (SIAM) followed suit in 1956 with support from the NSF. Due to the retrospective American translation program, the early results of the Russian school of the 1930s in non-linear mechanics and dynamical systems around Nikolay M. Krylov and Nikolay Bogolyubov became internationally better known. These results had built on even earlier work both by Henri Poincaré and Aleksandr Lyapunov at the turn of the century and connected to research by Russian physicists such as Leonid Mandelstam and by industrial mathematicians such as the Dutch radio-engineer Balthasar van der Pol. Some indigenous traditions (George Birkhoff as a follower of Poincaré), but above all the presence in the U.S. of early immigrants such as Minorsky (see above) from Russia with interests in applications ensured that the Russian results did not fall on totally unprepared ground in the West.

The 'Sputnik crisis' in 1957 caused American mathematicians to look even more closely at the work being done in the Soviet Union. However, the role of English as the lingua franca in the sciences and in mathematics would soon become overwhelming. At the ICM in Nice in 1970 all the plenary speakers, including the Russians, gave their talks in English with the exception of Lev Pontryagin, who used French.

Inner-German mathematical communication seems to have helped in overcoming the language barrier between Russians and Western mathematicians. Although the East and West Germans had each officially belonged to their own adherent organization of the IMU from 1964 – and their relationship has therefore to be considered 'international' in the understanding of this presentation – they continued to collaborate in projects such as editing the leading abstract journal *Zentralblatt für Mathematik* until the 1970s. Many reviews of

[34] One East German (E. Hölder) and one Pole (A. Mostowski) are listed in the Proceedings as belonging to the German and Polish delegations. However, neither of them appear as "members" or "authors" of the Congress and they were probably not present, maybe due to visa restrictions from either their own countries or the American side. The politically motivated visa problems which Laurent Schwartz faced before he could participate in the congress and receive the Fields medal there have been described in his autobiography (1997).

Russian papers were written by East Germans who knew the language from school or had even studied mathematics in the Soviet Union [37].

The economic superiority of the West, which gradually began to affect the infrastructure of mathematical research too and perpetuated the brain drain of mathematicians from many countries worldwide in particular to the U.S., increasingly defined the rules in scientific and cultural communication. Of course there continued and continue to exist until today many mathematicians in Europe and other places in the world, who work on an equal level with the Americans, among them many Russians. Nevertheless there is no doubt about the superior technological and industrial infra-structure, in particular with respect to computing facilities and software-development, which existed in the U.S., even long before the current revolution in information technology. Although this superiority was sometimes opposed with resentment (documented for instance by the British mathematician Alan Turing), it was admitted even self-critically by Russians[35] and by Western European applied mathematicians such as the Frenchman Louis Couffignal, the cybernetics pioneer, and by Jacques-Louis Lions, the numerical analyst. However, the close collaboration of Lions with Soviet applied mathematician Guriy Marchuk in the 1960s showed that the assumption of underdevelopment and isolation of Russian computing does not give the full picture. Lions sometimes felt that the lack of instruments increased the theoretical depth of their collaborative work.

Until the end of the Cold War, international relations in mathematics, at least in Western countries, were very much characterized by movement of personnel and human resources. Shortly before the Iron Curtain came down, Duren, said about "foreign graduate students":

> *Besides the women, the other unanticipated source of mathematical talent that made the crops of expansion Ph.D.s after 1963 better than we had any right to expect came from abroad. Their numbers have been increasing year by year, relative to native-born Americans, until in 1987 more than half of American Ph.D. degrees in mathematics were awarded to foreign students. . . . These students are not only selected for ability from a world pool of mathematical talent (excluding only the Soviet Bloc countries), they also tend to be better trained in certain areas such as hard analysis and mechanics. This may make them better than Americans in applied mathematics.* [14, p. 436]

Duren cites also the former French cabinet member J. J. Servan Schreiber, who insisted at the same time (1987) that "America must remain the world's graduate university for the sake of both U.S. and world economic, technological, and intellectual development." [14, p. 437]

Problems of international communication between East and West concerned not just languages and economic infrastructure but remained very much political until the end of the Cold War and the fall of the Iron Curtain, including problems of military funding, which often caused discussions within national communities of mathematicians.

It was only after the political turn around 1990 that historical reports appeared regularly in journals such as the *Notices of the American Mathematical Society* about former travel-restrictions for East European mathematicians, special programs such as IREX (International Research and Exchanges Board), which had allowed some exchange of personnel, defections of some scholars to the West etc. The complete history of these abnormalities and disturbances of international communication has yet to be written.

[35] See the contribution of A.P. Ershov and M. R. Shura-Bura in [23], written long before the fall of the Iron Curtain.

Lehto describes in detail the controversies and diplomatic efforts around the 1983 ICM in Warsaw (postponed due to martial law there) [20]. He discusses the negotiations about the membership of the People's Republic of China in IMU. He suggests, based on the experiences of the IMU representatives who negotiated in Moscow 1980, that the anti-Semitism which disturbed the Russian relationship with the IMU, was not necessarily imposed by the political regime but was supported by influential Russian mathematicians themselves [20, p. 217]. However, there is little doubt that the suppression of Jewish mathematicians in the Soviet Union, denying them travel to the West, often combined with attacks on U.S.-supported Israel, reflected a growing feeling among leading figures of the system of being doomed in the Cold War, and anti-Semitism thus foreshadowed the fall of the Iron Curtain. Indeed, since the 1970s there had been an emigration of Russian-Jewish mathematicians, mostly via Israel, a movement which overlaps with the third and most recent period of international mathematical communication to be discussed in this paper.

4. International mathematical communication after the fall of the Iron Curtain and conclusions

The Iron Curtain fell in 1989 not least due to Western superiority in communication technologies, partly based on mathematics. Because that singular political event entailed another wave of worldwide migration of mathematicians one may safely date the last and most recent period in international mathematical communication from that year 1989, epitomizing an overlap of deep political and technological changes. It is probably too early to come to final conclusions about these very recent events and their consequences for world mathematics. Therefore we will try to weave several of them – with some emphasis on variants and invariants – into a tentative summary of a century of international mathematical communication.

International mathematical collaboration on the research level has continued to celebrate success: two of the most spectacular recent mathematical accomplishments, the proofs of the Fermat and the Poincaré conjectures have profited much – partly when the proofs were checked for correctness – from the internationalism of mathematicians.

There are obvious recent changes in communication technologies, such as the posting of articles on the arXiv.org website starting in August 1991, which is now hosted and operated by Cornell University in the United States and continues and replaces the tradition of pre-prints from the pre-computer age. Old and new problems of publishing peer reviewed articles have been hotly debated recently on an international scale, for instance in responses to and in massive support for an initiative by Fields medalist Timothy Gowers. The new element caused by modern technologies is the shift of the working load in the publishing process to scholars and academic institutions and the diminished role of print on paper, with undiminished or even increased profits on the part of commercial publishers. The resulting conflict reignites old tensions between the mathematical community and commercial publishers dating back to the 1920s. This older tradition has also some potential to dampen the current crisis as revealed in the following passage from the 2012-memo "The Cost of Knowledge":[36]

[36] http://thecostofknowledge.com/. Cf. statement of purpose, p. 3.

> *One reason for focusing on Elsevier rather than, say, Springer is that Springer has had a rich and productive history with the mathematical community.*

Not only in relation to publishers but also among mathematicians themselves, the ethics of professional competition and publishing has been increasingly discussed in recent decades, triggered by spectacular events such as Grigoriy Perelman's refusal to accept the Clay Institute Millenium prize money for his confirmation of the Poincaré conjecture.

New forms of collaboration between mathematicians, based on the new communication technologies, have been proposed and enacted, for example the Polymath Project, initiated by Gowers in 2009.

Discussions on and comparison of the various national educational systems from the primary up to the tertiary level continue unabated. Topics such as mathematical competence of teachers and the relation between authority and freedom in class room remain on the agenda. It seems as though the strong traditions of school training in the sciences and mathematics in East European countries have been jeopardized and partly destroyed after the fall of the Iron Curtain. The usually higher scoring of pupils from some Asian countries in comparative surveys may indicate a further rise of mathematics in Asia in the future. The International Commission for Mathematical Instruction (ICMI), founded at the ICM in Rome 1908, has meanwhile expanded its activities considerably and organizes independent international congresses (ICME since 1969). While historiography of mathematics has often been recognized in principle as an important part and stimulus of mathematical education, as for example in the initiative at Swarthmore in 1945 discussed above, this recognition has not necessarily translated into an emphasis on history in national educational programs. A criticism of the neglect of the history of mathematics, within a representative publication of the mathematics education community, has recently been published by one educator [17]. In this respect the situation seems to have deteriorated in the last two decades, again partly as a result of the dissolution of strong centers of history of mathematics in Eastern Europe [46].

Throughout the past century the differences in educational systems and in research priorities have been a stimulus for world-wide collaboration. Even today there seem to exist certain advantages in Europe and Asia in some fields of research and school education on which the U.S. continue to rely. In a 1998 report of the American NSF one finds the following evaluation:

> *Although the United States is the strongest national community in the mathematical sciences, this strength is somewhat fragile. If one took into account only home-grown experts, the United States would be weaker than Western Europe. ... Western Europe is nearly as strong in mathematics as the United States, and leads in important areas. It has also benefited by the presence of émigré Soviet mathematical scientists.*[37]

Indeed, on the side of the mobility of personnel, the most visible change of international mathematical communication in recent decades is the massive emigration of Russian mathematicians as described by the Israeli-Russian functional analyst Vitali Milman, who himself was instrumental in this process:

> *The emigration of mathematicians from the Soviet Union to Israel began in the early 1970s. ... Every mathematical center in the West was touched and en-*

[37] *Report of the Senior Assessment Panel for the International Assessment of the U.S. Mathematical Sciences*, March 1998, 69 pp., p. 27, at http://www.nsf.gov/pubs/1998/nsf9895/nsf9895.pdf.

> *riched by this movement. But only a few people understood that, while beneficial for these individual centers, it bore elements of tragedy for mathematics as a whole.* [24, p. 216]

Milman goes on describing the "elements of tragedy" and the losses for Russian and World mathematics induced by this process, the importance of national mentalities in the creation of mathematics, referring to the

> *The concept of the 'Russian mathematical school', . . . which is extremely difficult to explain to a Westerner, encompasses traditions that prescribe ways of studying mathematics and a code of behavior for mathematicians. It is more an intellectual necessity (and a game) than it is work. Scholars raised in the traditions of the Russian mathematical school do not study mathematics for the sake of a salary.* [24, p. 216]

While Milman arrives at an optimistic conclusion and assumes that "the Russian mathematical school and its traditions will be preserved: they will take root in a new country and a new environment." [24, p. 228], his Russian colleague Anatoliy Vitushkin, publishing in the same volume, is less upbeat:

> *Perestroika has brought a lot of changes: one can go anywhere, [however] those who work in state-controlled institutes earn ridiculously small salaries. . . . Many mathematicians have left for other countries. . . . They appear to lose shape from hunting for jobs. Not all of them, certainly. Manin is always Manin, and Arnold as well . . .* . [50, p. 473]

Similar and broader concerns have been discussed in recent years on the pages of the *Notices of the American Mathematical Society*. Fears were articulated for the education of young Russian mathematicians who had been traced and recruited for instance through the system of mathematical 'Olympiads,' which had an international dimension as well. The prominent Belgian mathematician Pierre Deligne said earlier this year:

> *They have also the tradition of Olympiads, and they are very good at detecting promising people in mathematics early on in order to help them. The culture of seminars is in danger because it's important that the head of the seminars is working full-time in Moscow, and that is not always the case. There is a whole culture which I think is important to preserve. That is the reason why I used half of the Balzan Prize to try to help young Russian mathematicians.* [29, p. 183]

One Russian mathematician remembered the loss for those remaining in Russia, which was particularly strongly felt before the political turn:

> *Emigrants at that time disappeared completely behind the iron curtain, and we had a feeling that they were lost forever.* [30, pp. 164–165]

Together with Deligne, other mathematicians, particularly in the U.S., have helped to preserve the Russian mathematical culture and have learned from it, as exemplified by American support for the new 'Independent University of Moscow' [1].

The history of international mathematical communication and our discussion in particular have shown that international mathematical communication is not necessarily unproblematic or a guaranty for a healthy development of our science. 'Internationalization' (or 'internationality' considered as its result) without equal chances or even equal rights of the participants in international collaboration is generally problematic, as the extreme case of the Nazi strategies during the occupation of Europe show. The exodus of scientists from Europe was, to be sure, a source of a tremendous push in the 'internationalization' of mathematics, especially in the sense of *new* and literally unexpected personal encounters and oral communication. However, here, as in later examples of international mathematical communication mentioned in this paper, the historian cannot construct an uncritical success story but has also to look at the *losses* for mathematics and for its individuals, which were often equally substantial as the *gains*.

Acknowledgements. I am grateful to the School of Engineering and Applied Sciences of Harvard University for its hospitality during the writing of this paper. I am grateful to June Barrow-Green who proposed a considerable revision of the manuscript and improved the language as well.

References

[1] Anon., *In Appreciation of the AMS-fSU Aid Fund*, Notices of the American Mathematical Society **42** (1995), 476.

[2] Bell, E.T., *The Development of Mathematics*, 2nd ed. McGraw Hill, New York, 1945.

[3] Beaulieu, L., *Regard sur les mathématiques en France entre les deux guerres. Introduction*, Revue d'histoire des sciences **62** (2009), 9–38.

[4] Birkhoff, G., *Some Leaders in American Mathematics*, 1891–1941, in D. Tarwater (ed.), The Bicentennial Tribute to American Mathematics, 1776–1976. Mathematical Association of America, Washington, D.C., 1977, 25–78.

[5] Birkhoff, G.D., *Fifty Years of American Mathematics*, in American Mathematical Society Semicentennial Publications in Two Volumes, American Mathematical Society, NewYork, 1938, Vol. 2, 270–315.

[6] Booß-Bavnbek, B. and Høyrup, J. (eds.), *Mathematics and War*. Birkhäuser, Basel, 2003.

[7] Bru, B., *Souvenirs de Bologne*, Journal de la Société Française de Statistique **144** (2003), 135–226.

[8] Cartwright, M.L., *From non-linear oscillations to topological dynamics*, Journal of the London Mathematical Society **39** (1964), 193–201.

[9] Chandler, B., Magnus and W., *The History of Combinatorial Group Theory: A Case Study in the History of Ideas*, Springer, New York, 1982.

[10] Coifman, R.R. and Strichartz, R.S., *The School of Antoni Zygmund*, in P. Duren (ed.), A Century of Mathematics in America, 3 Parts, American Mathematical Society, Providence, RI, 1988–1989, Part 3, 343–368.

[11] Corry, L., *Modern Algebra and the Rise of Mathematical Structures*; second, revised edition, Birkhäuser, Basel, 2004.

[12] Curbera, G.P., *Mathematicians of the World, Unite! The International Congress of Mathematicians – a Human Endeavor*. A.K. Peters, Wellesley, 2009.

[13] Demidov, S., *Les relations mathématiques franco-russes entre les deux guerres mondiales*, Revue d'histoire des sciences **62** (2009), 119–142.

[14] Duren, W.L., *Mathematics in American Society 1888-1988. A Historical Commentary*, in P. Duren (ed.), A Century of Mathematics in America, 3 Parts, American Mathematical Society, Providence, RI, 1988–1989, Part 2, 399–447.

[15] Epple, M., Karachalios, A., and Remmert, V., *Aerodynamics and Mathematics in National Socialist Germany and Fascist Italy: A Comparison of Research Institutes*, Osiris, **20** (2005), 131–158.

[16] Gray, J. and Parshall, K.H. (eds.), *Episodes in the history of modern algebra (1800-1950)*, American Mathematical Society, Providence, 2007.

[17] Jankvist, U.T., *A century of mathematics education: ICMI's first hundred years*, Historia Mathematica **38** (2011), 292–302.

[18] Keyfitz, B.L., *Mathematics and industry: an interdisciplinary perspective*, in Chamizo, F., Quiros, A. (eds.) Madrid Intelligencer 2006. Springer, 2006, 22–26.

[19] Khesin, B. and Tabachnikov, S. (eds.), *Tribute to Vladimir Arnold*, Notices of the American Mathematical Society **59** (2012), 378–399.

[20] Lehto, O., *Mathematics Without Borders. A History of the International Mathematical Union*, Springer, New York, 1998.

[21] Mehrtens, H., *Ludwig Bieberbach and 'Deutsche Mathematik'*, in E. Phillips (ed.), Studies in the History of Mathematics. Mathematical Association of America, Washington, D.C., 1987, 195–241.

[22] ———, *Moderne, Sprache, Mathematik*, Suhrkamp, Frankfurt, 1990.

[23] Metropolis, N., J. Howlett, J., and Rota, G.-C. (eds.), *A History of Computing in the Twentieth Century: A Collection of Essays*, Academic Press, New York, 1980.

[24] Milman, V.D., *Observations on the Movement of People and Ideas in Twentieth-Century Mathematics*, in A.A. Bolibruch, Yu S. Osipov, Ya G. Sinai, (eds.), Mathematical Events of the Twentieth Century, Springer, Berlin, 2006, 215–241.

[25] Parshall, K.H., *Mathematics in National Contexts (1875–1900): An International Overview*, Proceedings of the International Congress of Mathematicians, Zürich 1994, Vol.2, Birkhäuser, Basel, 1995, 1581–1591.

[26] Pier, J.-P. (ed.), *Development of Mathematics 1900–1950*. Birkhäuser, Basel, 1954.

[27] _____ (ed.), *Development of Mathematics 1950–2000*. Birkhäuser, Basel, 2000.

[28] Price, D.J. de Solla, *Little Science, Big Science And Beyond*, Columbia University Press, New York, 1986.

[29] Raussen, M. and Skau, C., *Interview with Pierre Deligne*, Notices of the American Mathematical Society **61** (2014), 177–185.

[30] Retakh, V. (ed.), *Israel Moiseevich Gelfand. Part II*, Notices of the American Mathematical Society **60** (2013), 162–171.

[31] Richardson, R.G.D., *The Ph.D. Degree and Mathematical Research*, American Mathematical Monthly **43** (1936), 199–215.

[32] Rowe, D., *Disciplinary Cultures of Mathematical Productivity in Germany*, in V.R. Remmert, U.Schneider (eds.), Publikationsstrategien einer Disziplin—Mathematik in Kaiserreich und Weimarer Republik. Harrassowitz Verlag, Wiesbaden, 2008, 9–51.

[33] Saul, M., *Mathematics in a Small Place: Notes on the Mathematics of Romania and Bulgaria*, Notices of the American Mathematical Society **50** (2003), 561–565.

[34] Schappacher, N., *Seventy Years Ago: The Bourbaki Congress at El Escorial and Other Mathematical (Non) Events of 1936*, in F. Chamizo, A.Quiros, Madrid Intelligencer 2006. Springer, 2006, 8–15.

[35] Segal, S.L., *Mathematicians under the Nazis*, Princeton University Press, Princeton, NJ, 2003.

[36] Senechal, M., *The Continuing Silence of Bourbaki—An Interview with Pierre Cartier*, June 18, 1997, The Mathematical Intelligencer **20** (1998), no.1, 22–28.

[37] Siegmund-Schultze, R., *Dealing with the political past of East German mathematics*, The Mathematical Intelligencer **15** (1993), no.4, 27–36.

[38] _____, *"Scientific Control", in Mathematical Reviewing and German–U.S.–American relations between the Two World Wars*, Historia Mathematica **21** (1994), 306–329.

[39] _____, *National Styles in Mathematics between the World Wars?*, in E. Ausejo, M. Hormigon, M. (eds.), Paradigms and Mathematics, Siglo XXI de Espana Editores, Madrid, 1996, 243–253.

[40] _____, *The Emancipation of Mathematical Research Publishing in the United States from German Dominance (1878-1945)*, Historia Mathematica **24** (1997), 135–166.

[41] _____, *Rockefeller and the Internationalization of Mathematics Between the Two World Wars*, Birkhäuser, Basel, 2001.

[42] _____, *The Effects of Nazi Rule on the International Participation of German Mathematicians: An Overview and Two Case Studies*, in K. Parshall, A. Rice (eds.), Mathematics Unbound: The Evolution of an International Mathematical Research Community, 1800-1945. American Mathematical Society and London Mathematical Society, Providence and London, 2002, 335–357.

[43] _____, *The late arrival of academic applied mathematics in the United States: a paradox, theses, and literature*, N.T.M. International Journal of History and Ethics of Natural Sciences, Technology and Medicine (N.S.) **11** (2003), 116–127.

[44] _____, *Military Work in Mathematics 1914-1945: an Attempt at an International Perspective*, in B. Booß-Bavnbek, J. Høyrup (eds.), Mathematics and War. Birkhäuser, Basel, 2003, 23–82.

[45] _____, *Mathematicians fleeing from Nazi Germany: Individual Fates and Global Impact*, Princeton University Press, Princeton, 2009.

[46] _____, *Hans Wußing (1927–2011) and the blooming of the history of mathematics and sciences in the German Democratic Republic A biographical essay*, Historia Mathematica **39** (2012), 143–173.

[47] Tobies, R., *Biographisches Lexikon in Mathematik promovierter Personen an deutschen Universitäten und Technischen Hochschulen WS 1907/08 bis WS 1944/45*, Dr. Erwin Rauner Verlag, Augsburg, 2006.

[48] Ulam, S., *Adventures of a Mathematician*, Scribners, New York, 1976.

[49] Veblen, O., *George David Birkhoff(1884-1944)*, reprinted in *G.D. Birkhoff, Collected Mathematical Papers*, American Mathematical Society, New York, 1950, Vol.1, xv–xxi.

[50] Vitushkin, A.G., *Half a Century as One Day*, in A.A. Bolibruch, Yu S. Osipov, Ya G. Sinai (eds.), Mathematical Events of the Twentieth Century, Springer, Berlin, 2006, 449–473.

[51] Wavre, R., *The International Congresses of Mathematicians*, in F. Le Lionnais (ed.), Great Currents of Mathematical Thought, Dover, New York, 1971, Volume 1, 312–318. First French edition 1948.

[52] Weil, A., *History of Mathematics: Why and How?*, Proceedings of the International Congress of Mathematicians, Helsinki, 1978, Vol.1, Academia Scientiarum Fennica, Helsinki 1980, 227–236.

[53] Wilson, E.B., *Insidious Scientific Control*, Science **48** (1918), 491–493.

University of Agder Faculty of Engineering and Science, Gimlemoen Dept of Mathematics Post Box 422 NO-4604 Kristiansand, Norway

E-mail: reinhard.siegmund-schultze@uia.no

Mathematics of engineers: Elements for a new history of numerical analysis

Dominique Tournès

Abstract. The historiography of numerical analysis is still relatively poor. It does not take sufficient account of numerical and graphical methods created, used and taught by military and civil engineers in response to their specific needs, which are not always the same as those of mathematicians, astronomers and physicists. This paper presents some recent historical research that shows the interest it would be to examine more closely the mathematical practices of engineers and their interactions with other professional communities to better define the context of the emergence of numerical analysis as an autonomous discipline in the late 19th century.

Mathematics Subject Classification (2010). Primary 65-03; Secondary 01A55, 01A60.

Keywords. Mathematics of engineers, numerical analysis, nomography, civil engineering, topography, ballistics, hydraulics, linear systems, differential equations, dynamical systems.

1. Introduction

Few recent books have been devoted to the history of numerical analysis. Goldstine [18] was a pioneer. His work focuses primarily on identifying numerical methods encountered in the works of some great mathematicians: Newton, Maclaurin, Euler, Lagrange, Laplace, Legendre, Gauss, Cauchy and Hermite. The main problems are the construction of logarithmic and trigonometric tables necessary for astronomical calculations, Kepler's equation, the Lunar theory and its connection with the calculation of longitudes, the three-body problem and, more generally, the study of perturbations of orbits of planets and comets. Through these problems we assist to the birth of finite difference methods for interpolating functions and calculating quadratures, developments in series or continued fractions for solving algebraic equations and differential equations, and the method of least squares for finding optimal solutions of linear systems with more equations or less equations than unknowns. At the end of the book, a few pages involve Runge, Heun, Kutta, Moulton, that is to say, some characters who can be considered as being the first applied mathematicians identified as such in the late 19th century and the beginning of the 20th. In Goldstine's survey, numerical analysis is thus the fruit of a few great mathematicians who developed the foundations of today's numerical methods by solving some major problems of astronomy, celestial mechanics and rational mechanics. These numerical methods were then deepened by professional applied mathematicians appearing in the late 19th century, which was the time when numerical analysis, as we know it today, structured itself into an autonomous discipline. In this story, a few areas

Proceedings of the International Congress of Mathematicians, Seoul, 2014

of inspiration and intervention other than astronomy are sometimes mentioned incidentally, but no engineer is explicitly quoted.

While Goldstine actually begins his history in the 16th century, Chabert [7] gives more depth to the subject by examining numerical algorithms in a variety of texts from various civilizations since Antiquity. Besides the famous previously mentioned problems of astronomy such as Kepler's equation, the determination of orbits of comets, the brightness of stars, etc., there are some references to other domains, for example the theory of vibrating strings or the signal theory. Some engineers are mentioned, in general in connection with secondary points. Only one of them, Cholesky, is quoted for a significant contribution consisting in an original method for solving linear systems (see Section 3). Despite these few openings compared to previous work, most numerical analysis questions addressed in Chabert's book are presented as abstract mathematical problems, out of context.

In a more recent collective book edited by Bultheel and Cools [6], the birth of modern numerical analysis is located precisely in 1947, in a paper of John von Neumann (1903–1957) and Herman Goldstine (1913–2004) [23] which analyzes for the first time in detail the propagation or errors when solving a linear system, in conjunction with the first uses of digital computers. The authors recognize naturally that a lot of numerical calculations were made long before this date in various questions of physics and engineering, but for them the problem of the practical management of calculations made by computer actually founds the field of numerical analysis and this apparently technical problem is at the origin of the considerable theoretical developments that this domain generated since the mid-20th century. In this book written not by historians but by specialists of numerical analysis, it is interesting to note that the accepted actors of the domain do not trace the history of their discipline beyond what characterizes their current personal practices.

In fact, the birth of numerical analysis, in the modern sense of the term, should not be connected to the advent of digital computers, but to the distinction between pure mathematics and applied mathematics (formerly "mixed mathematics"), which is clarified gradually throughout the 19th century with a more and more marked separation between the two domains in scientific journals, institutions and university positions[1]. The development of new calculating instruments – before computers, there were numerical and graphical tables, slide rules, mechanical instruments of integration, desktop calculators, etc. – has also contributed to set up a new equilibrium between analytical, numerical and graphical methods. This is actually around 1900 that mathematicians began to formulate, in concrete terms, what is meant by "applied mathematics". Germany, and particularly Göttingen, played a leading role in this international process of institutionalization of applied mathematics as an autonomous domain [26, p. 60–63]. Encouraged by Felix Klein, Carl Runge (1856–1927) and Rudolf Mehmke (1857–1944) assumed in 1901 the editorship or the *Zeitschrift für Mathematik und Physik* and devoted this journal to applied mathematics. In 1904, Runge accepted the first full professorship of applied mathematics at the University of Göttingen. In 1907, German applied mathematicians adopted the following definition:

> The essence of applied mathematics lies in the development of methods that will lead to the numerical and graphical solution of mathematical problems.[2]

[1] A very interesting workshop on this subject took place in March 2013 in Oberwolfach, organized by Moritz Epple, Tinne Hoff Kjeldsen and Reinhard Siegmund-Schultze, and entitled "From 'Mixed' to 'Applied' Mathematics: Tracing an important dimension of mathematics and its history" [13].

[2] "Das Wesen der angewandten Mathematik liegt in der Ausbildung und Ausübung von Methoden zur numerischen und graphischen Durchführung mathematischer Probleme" (quoted in [27, p. 724]).

Recent research has shown that engineers have constituted a bridge between mathematics and their applications since the 18th century, and that problems encountered in ballistics, strength of materials, hydrodynamics, steam engines, electricity and telephone networks played also an important role in the creation of original numerical and graphical methods of computation. In fact, the mathematical needs of engineers seem very different from those of mathematicians. To illustrate this with a significant example, consider the problem of the numerical solution of equations, a pervasive problem in all areas of mathematics intervention. Léon-Louis Lalanne (1811–1892), a French civil engineer who, throughout his career, sought to develop practical methods for solving equations, wrote what follows as a summary when he became director of the École des ponts et chaussées:

> The applications have been, until now, the stumbling block of all the methods devised for solving numerical equations, not that, nor the rigor of these processes, nor the beauty of the considerations on which they are based, could have been challenged, but finally it must be recognized that, while continuing to earn the admiration of geometers, the discoveries of Lagrange, Cauchy, Fourier, Sturm, Hermite, etc., did not always provide easily practicable means for the determination of the roots.[3]

Lalanne says that as politely as possible, but his conclusion is clear: the methods advocated by mathematicians are not satisfactory. These methods are complicated to understand, long to implement and sometimes totally impracticable for ground engineers, foremen and technicians, who, moreover, did not always receive a high-level mathematical training.

Given such a situation, 19th century engineers were often forced to imagine by themselves the operational methods and the calculation tools that mathematicians could not provide them. The objectives of the engineer are not the same as those of the mathematician, the physicist or the astronomer: the engineer rarely needs high accuracy in his calculations, he is rather sensible to the speed and simplicity of their implementation, especially since he has often to perform numerous and repetitive operations. He needs also methods adapted for use on the ground, and not just for use at the office. Finally, priority is given to methods that avoid performing calculations by oneself, methods that provide directly the desired result through a simple reading of a number on a numerical or graphical table, on a diagram, on a curve or on the dial of a mechanical instrument.

In this paper, I would want to show, through some examples from recent historical research, that the engineers, so little mentioned so far in the historiography of numerical analysis, have contributed significantly throughout the 19th century to the creation of those numerical and graphical methods that became an autonomous discipline around 1900. More than that, I shall underline that their practical methods have been sometimes at the origin of new theoretical problems that inspired also pure mathematicians.

[3] "Les applications ont été, jusqu'à ce jour, la pierre d'achoppement de tous les procédés imaginés pour la résolution des équations numériques, non pas que, ni la rigueur de ces procédés, ni la beauté des considérations sur lesquelles ils se fondent, en aient reçu la moindre atteinte; mais enfin il bien reconnaître que, sans cesser de mériter l'admiration des géomètres, les découvertes de Lagrange, de Cauchy, de Fourier, de Sturm, d'Hermite, etc., n'ont pas fourni toujours des moyens facilement praticables pour la détermination des racines" [20, p. 1487].

2. From Civil engineering to nomography

The 19th century is the moment of the first industrial revolution, which spreads throughout the Western world at different rates in different countries. Industrialization causes profound transformations of society. In this process, the engineering world acquires a new identity, marked by its implications in the economic development of industrial states and the structuration of new professional relationships that transcend national boundaries. Linked to the industrial revolution, enormous computational requirements appeared during the 19th century in all areas of engineering sciences and caused an increasing mathematization of these sciences. This led naturally to the question of engineering education: how were engineers prepared to use high-level mathematics in their daily work and, if necessary, to create by themselves new mathematical tools?

The French model of engineering education in the early 19th century is that of the École polytechnique, founded in 1794.[4] Although it had initially the ambition to be comprehensive and practice-oriented, this school promoted quickly a high-level teaching dominated by mathematical analysis. This theoretical teaching was then completed, from the professional point of view, by two years in application schools with civil and military purposes. Such a training model, which subordinates practice to theory, has produced a corporation of "scholarly engineers" capable of using the theoretical resources acquired during their studies to achieve an unprecedented mathematization of the engineering art. This model is considered to have influenced the creation of many polytechnic institutes throughout Europe and to the United States.

A paradigmatic example of a corpus of mathematical tools, constituting an autonomous knowledge which was created from scratch by engineers themselves to meet their needs, is that of nomography.[5] The main purpose of nomography is to construct graphical tables to represent any relationship between three variables, and, more generally, relationships between any number of variables. Among the "Founding Fathers" of nomography, four were students at the École polytechnique: Lalanne, Charles Lallemand (1857–1938), Maurice d'Ocagne (1862–1938) and Rodolphe Soreau (1865–1936). The only exception in this list is the Belgian engineer Junius Massau (1852–1909), an ancient student and then professor at the school of civil engineering of the University of Ghent, but, in this school of civil engineering, the training was comparable to that of the École polytechnique, with high-level courses of mathematics and mechanics.

During the years 1830–1860, the sector of public works experiences a boom in France and more generally in Europe. The territories of the different countries are covered progressively by vast networks of roads, canals, and, after 1842, of railways. These achievements require many tedious calculations of surfaces of "cut and fill" on cross-sections of the ground. Cut and fill is the process of earthmoving needed to construct a road, a canal or a railway. You have to cut land where the ground level is too high and then transport this land to fill the places where the ground level is too low. And to calculate roughly the volume of land to be transported, you have to decompose this volume in thin vertical slices, evaluate the area of each slice and sum all these elementary areas.

Civil engineers tried different methods of calculation more or less expeditious. Some, like Gaspard-Gustave Coriolis (1792–1843), have calculated numerical tables giving the sur-

[4] On the professional milieu of French engineers during the 19th century and the École polytechnique, see the papers by Bruno Belhoste and Konstantinos Chatzis ([2, 9]).

[5] This Section is an abridged and synthetic version of developments contained in my papers [30, 32, 34].

faces directly based on a number of features of the road and its environment. Other engineers, especially in Germany and Switzerland, designed and built several kinds of planimeters, that is mechanical instruments used to quickly calculate the area of any plane surface. These planimeters, which concretize the continuous summation of infinitesimal surfaces, had significant applications in many other scientific fields beyond cuts and fills. Still others, like Lalanne, have imagined replacing numerical tables by graphical tables, cheaper and easier to use. It is within this framework that nomography developed itself and was deepened throughout the second half of the 19th century.

First principles of nomography. The departure point of nomography lies in the fact that a relationship between three variables α, β and γ can be considered, under certain conditions, as the result of the elimination of two auxiliary variables x and y between three equations, each containing only one of the initial variables. One can then represent the equation by three sets of lines in the plane x-y, one of them parametrized by α, the second by β and the third by γ. On this kind of graphical table, called a "concurrent-line abaque", a solution of the equation corresponds to an intersection point of three lines.

Isolated examples of graphical translation of double-entry tables are found already in the first half of the 19th century, mainly in the scope of artillery, but this is especially Lalanne who gave a decisive impetus to the theory of graphical tables. In 1843, he provided consistent evidence that any law linking three variables can be graphed in the same manner as a topographic surface using its marked level lines. His ideas came to a favorable moment. Indeed, the Act of June 11, 1842 had decided to establish a network of major railway lines arranged in a star from Paris. To run the decision quickly, one felt the need for new ways of evaluating the considerable earthworks to be carried out. In 1843, the French government sent to all engineers involved in this task a set of graphical tables for calculating the areas of cut and fill on the profile of railways and roads.

Curves other than straight lines are difficult to construct on paper. For this reason, Lalanne imagined the use of non-regular scales on the axes for transforming curves into straight lines. By analogy with the well-known optical phenomenon previously used by certain painters, he called "anamorphosis" this general transformation process. After Lalanne, the graphical tables resting on the principle of concurrent lines spread rapidly until becoming, in the third quarter of the 19th century, very common tools in the world of French engineers.

Massau succeeded Lalanne to enrich the method and its scope of application. For that, he introduced a notion of generalized anamorphosis, seeking what are the functions that can be represented using three pencils of lines. Massau put in evidence that a given relationship between three variables can be represented by a concurrent-straight-line abaque if, and only if, it can be put into the form of a determinant of the type

$$\begin{vmatrix} f_1(\alpha) & f_2(\alpha) & f_3(\alpha) \\ g_1(\beta) & g_2(\beta) & g_3(\beta) \\ h_1(\gamma) & h_2(\gamma) & h_3(\gamma) \end{vmatrix} = 0.$$

These determinants, called "Massau determinants", played an important role in the subsequent history of nomography; they are encountered in research until today. As an application of this new theory, Massau succeeded in simplifying Lalanne's abaques for cuts and fills. With Massau's publications, the theory of abaques was entering into a mature phase, but in the same time a new character intervened to orient this theory towards a new direction.

From concurrent-line abaques to alignment nomograms. In 1884, when he is only 22 years old, d'Ocagne observes that most of the equations encountered in practice can be represented by an abaque with three systems of straight lines and that three of these lines, each taken in one system, correspond when they meet into a point. His basic idea is then to construct by duality, by substituting the use of tangential coordinates to that of punctual coordinates, a figure in correlation with the previous one: each line of the initial chart is thus transformed into a point, and three concurrent lines are transformed into three aligned points. The three systems of marked straight lines become three marked curves. Through this transformation, a concurrent-straight-line abaque becomes an "alignment abaque", which is easier to use.

A given relationship between three variables is representable by an alignment abaque if, and only if, it can be put into the form of a Massau determinant, because it is clear that the problem of the concurrency of three straight lines and the problem of the alignment of three points, dual to each other, are mathematically equivalent. As his predecessors, d'Ocagne applied immediately his new ideas to the problem of cuts and fills, actually one of the main problems of civil engineering.

After this first achievement in 1891, d'Ocagne deepened the theory and applications of alignment abaques until the publication of a large treatise in 1899, the famous *Traité de nomographie*, which became for a long time the reference book of the new discipline. A little later, he introduced the generic term "nomogram" to replace "abaque", and the science of graphical tables became "nomography". From there, alignment nomograms were quickly adopted by many engineers for the benefit of the most diverse applications. At the turn of the 20th century, nomography was already an autonomous discipline well established in the landscape of applied sciences.

Mathematical implications of nomography. The mathematical practices of engineers are often identified only as "applications", which is equivalent to consider them as independent from the development of mathematical knowledge in itself. In this perspective, the engineer is not supposed to develop a truly mathematical activity. We want to show, through the example of nomography, that this representation is somewhat erroneous: it is easy to realize that the engineer is sometimes a creator of new mathematics, and, in addition, that some of the problems which he arises can in turn irrigate the theoretical research of mathematicians.

Firstly, the problem of general anamorphosis, that is to say, of characterizing the relationships between three variables that can be put in the form of a Massau determinant, has inspired many theoretical research to mathematicians and engineers: Cauchy, Saint-Robert, Massau, Lecornu, and Duporcq have brought partial responses to this problem before that in 1912 the Swedish mathematician Thomas Hakon Gronwall (1877–1932) gives a complete solution resulting in the existence of a common integral to two very complicated partial differential equations. But, as one can easily imagine, this solution was totally inefficient, except in very simple cases.

After Gronwall, other mathematicians considered the problem of anamorphosis in a different way, with a more algebraic approach that led to study the important theoretical problem of linear independence of functions of several variables. These mathematicians, like Kellogg in the US, wanted to find a more practical solution not involving partial differential equations. A complete and satisfactory solution was finally found by the Polish mathematician Mieczyslaw Warmus (1918–2007). In his Dissertation of 1958, Warmus defined precisely what is a nomographic function, that is a function of two variables that can be

represented by an alignment nomogram, and classified nomographic functions through homography into 17 equivalence classes of Massau determinants. Moreover, he gave an effective algorithm for determining if a function is nomographic and, if true, for representing it explicitly as a Massau determinant.

Beyond the central problem of nomographic representation of relationships between three variables, which define implicit functions of two variables, there is the more general problem of the representation of functions of three or more variables. Engineers have explored various ways in this direction, the first consisting in decomposing a function of any number of variables into a finite sequence of functions of two variables, which results in the combined use of several nomograms with three variables, each connected to the next by means of a common variable.

Such a practical concern was echoed unexpectedly in the formulation of the Hilbert's 13th problem, one of the famous 23 problems that were presented at the International Congress of Mathematicians in 1900 [19]. The issue, entitled "Impossibility of the solution of the general equation of the 7th degree by means of functions of only two arguments" is based on the initial observation that up to the sixth degree, algebraic equations are nomographiable.

Indeed, up to the fourth degree, the solutions are expressed by a finite combination of additions, subtractions, multiplications, divisions, square roots extractions and cube roots extractions, that is to say, by functions of one or two variables. For the degrees 5 and 6, the classical Tschirnhaus transformations lead to reduced equations whose solutions depend again on one or two parameters only. The seventh degree is then the first actual problem, as Hilbert remarks:

> Now it is probable that the root of the equation of the seventh degree is a function of its coefficients which does not belong to this class of functions capable of nomographic construction, i. e., that it cannot be constructed by a finite number of insertions of functions of two arguments. In order to prove this, the proof would be necessary that the equation of the seventh degree is not solvable with the help of any continuous functions of only two arguments [19, p. 462].

In 1901, d'Ocagne had found a way to represent the equation of the seventh degree by a nomogram involving an alignment of three points, two being carried by simple scales and the third by a double scale. Hilbert rejected this solution because it involved a mobile element. Without going into details, we will retain that there has been an interesting dialogue between an engineer and a mathematician reasoning in two different perspectives. In the terms formulated by Hilbert, it was only in 1957 that the 13th problem is solved negatively by Vladimir Arnold (1937–2010), who proved to everyone's surprise that every continuous function of three variables could be decomposed into continuous functions of two variables only.

3. From topography to linear systems

The French military engineer André-Louis Cholesky (1875–1918) offers us the occasion of a perfect case study. Before 1995, not many details were known on his life. In 1995 (120 years after his birth), the documents about him kept in the archives of the army at the Fort de Vincennes (near Paris) were open to the public. In 2003, we had the chance that a grandson of Cholesky, Michel Gross, donated the personal archives of his grandfather to the École

polytechnique.[6]

Cholesky was born on 15 October 1875, in Montguyon, a village near Bordeaux, in the south-west of France. On October 1895, he was admitted to the École polytechnique and, two years later, he was admitted as a *sous-lieutenant* at the École d'application de l'artillerie et du génie in Fontainebleau. He had to spend one year at the school and then to serve for one year in a regiment of the army. He had there courses on artillery, fortification, construction, mechanics, topography, etc.

Cholesky as a topographer. Between 1902 and 1906, he was sent to Tunisia and then to Algeria for several missions. In 1905, he was assigned to the Geographical Service of the Staff of the Army. In this service, there were a section of geodesy and a section of topography. Around 1900, following the revision of the meridian of Paris, the extent of the meridian of Lyon and a new cadastral triangulation of France had been decided. These missions were assigned to the section of geodesy together with the establishment of the map of Algeria, and a precise geometric levelling of this country. The problem of the adjustment (or compensation) of networks (corrections to be brought to the measured angles) concerned many officers of the Geographical Service, eager to find a simple, fast and accurate method. According to Commandant Benoît, one of his colleagues, it was at this occasion that Cholesky imagined his method for solving the equations of conditions by the method of least squares.

Cholesky is representative of these "scholarly engineers" of whom we spoke above. Due to his high-level mathematical training, he was able to work with efficiency and creativity in three domains: as a military engineer, specialized in artillery and in topography, able to improve and optimize the methods used on the ground at this time; as a mathematician able to create new algorithms when it is necessary; and as a teacher (because in parallel to his military activities, he participated during four years to the teaching by correspondence promoted by the École spéciale des travaux publics founded in Paris by Léon Eyrolles).

Concerning topography, Cholesky is well known among topographers for a levelling method of his own: the method of double-run levelling (*double cheminement* in French). Levelling consists in measuring the elevation of points with respect to a surface taken as reference. This surface is often the geoid in order to be able to draw level curves, also called "contour lines". Double-run levelling consists in conducting simultaneously two separate survey traverses, very close to each other, and comparing the results so as to limit the effects of some instrumental defects. This method is still taught and used today.

Cholesky's method for linear systems. As said before, Cholesky is a good example of an engineer creating a new mathematical method and a new algorithm of calculation for his own needs. Cholesky's method for linear systems is actually an important step in the history of numerical analysis. A system of linear equations has infinitely many solutions when the number of unknowns is greater than the number of equations. Among all possible solutions, one look for the solution minimizing the sum of the squares of the unknowns. This is the case in the compensation of triangles in topography which interested Cholesky. The method of least squares is very useful and is much used in many branches of applied mathematics (geodesy, astronomy, statistics, etc.) for the treatment of experimental data and fitting a mathematical model to them. This method was published for the first time by Legendre in 1806. Its interpretation as a statistical procedure was given by Gauss in 1809.

[6]Claude Brezinski has classified these archives and published a lot of papers about the life and work of Cholesky: see [3], [4] and [5]. For this Section, I took a lot of information in these papers.

As it is known, the least square method leads to a system with a symmetric positive definite matrix. Let us describe Cholesky's method to solve such a system. Let A be a symmetric positive definite matrix. It can be decomposed as $A = LL^T$, where L is a lower triangular matrix with positive diagonal elements, which are computed by an explicit algorithm. Then the system $Ax = b$ can be written as $LL^T x = b$. Setting $y = L^T x$, we have $Ly = b$. Solving this lower triangular system gives the vector y. Then x is obtained as the solution of the upper triangular system $L^T x = y$.

What was the situation before Cholesky? When the matrix A is symmetric, Gauss method makes no use of this property and needs too many arithmetical operations. In 1907, Otto Toeplitz showed that an Hermitian matrix can be factorized into a product LL^* with L lower triangular, but he gave no rule for obtaining the matrix L. That is preceisely what Cholesky did in 1910. Cholesky's method was presented for the first time in 1924 in a note published in the *Bulletin géodésique* by commandant Benoît, a French geodesist who knew Cholesky well, but the method remained unknown outside the circle of French military topographers. Cholesky method was rebirth by John Todd who taught it in his numerical analysis course at King's College in London in 1946 and thus made it known. When Claude Brezinski classified Cholesky's papers in 2003, he discovered the original unpublished manuscript where Cholesky explained his method[7]. The manuscript of 8 pages is dated 2 December 1910. That was an important discovery for the history of numerical analysis.

4. From ballistics to differential equations

The main problem of exterior ballistics is to determine the trajectory of a projectile launched from a cannon with a given angle and a given velocity. The differential equation of motion involves the gravity g, the velocity v and the tangent inclination θ of the projectile, and the air resistance $F(v)$, which is an unknown function of v:[8]

$$g\,\mathrm{d}(v\cos\theta) = vF(v)\,\mathrm{d}\theta.$$

To calculate their firing tables and to adjust their cannons, the artillerymen have used for a long time the assumption that the trajectory is parabolic, but this was not in agreement with the experiments. Newton was the first to research this topic taking into account the air resistance. In his *Principia* of 1687, he solved the problem with the hypothesis of a resistance proportional to the velocity, and he got quite rough approximations when the resistance is proportional to the square of the velocity. After Newton, Jean Bernoulli discovered the general solution in the case of a resistance proportional to any power of the velocity, but his solution, published in the *Acta Eruditorum* of 1719, was not convenient for numerical computation.

This problem of determining the ballistic trajectory for a given law of air resistance is particularly interesting because it stands at the crossroads of two partly contradictory concerns: on the one hand, the integration of the differential equation of motion is a difficult

[7]This manuscript has been published in 2005 in the *Revue d'histoire des mathématiques* [3].

[8]In fact, the problem is more complex because we must take into account other factors like the variations of the atmospheric pressure and temperature, the rotation of the Earth, the wind, the geometric form of the projectile and its rotation around its axis, etc. However these effects could be often neglected in the period considered here, because the velocities of projectiles remained small.

problem which interests the mathematicians from the point of view of pure analysis; on the other hand, the artillerymen on the battlefield must determine quickly the firing angle and the initial velocity of their projectile in order to attain a given target, and for that practical purpose they need firing tables precise and easy to use. This tension between theoreticians, generally called ballisticians, and practitioners, described rather to be artillerymen, is seen in all synthesis treatises of the late 19th and early 20th century. I shall content myself with one quotation to illustrate this tension. In 1892, in the French augmented edition of his main treatise, Francesco Siacci (1839–1907), a major figure in Italian ballistics, writes:

> Our intention is not to present a treatise of pure science, but a book of immediate usefulness. Few years ago ballistics was still considered by the artillerymen and not without reason as a luxury science, reserved for the theoreticians. We tried to make it practical, adapted to solve fast the firing questions, as exactly as possible, with economy of time and money.[9]

By these words, Siacci condemns a certain type of theoretical research as a luxury, but he condemns also a certain type of experimental research that accumulates numerous and expensive firings and measurements without obtaining convincing results.

Of course, the problem of integrating the ballistic equation is difficult. Many, many attempts have been done to treat this equation mathematically with the final objective of constructing firing tables. We can organize these attempts throughout two main strategies, one analytical and one numerical.

Analytical approach of the ballistic differential equation. The analytical strategy consists in integrating the differential equation in finite terms or, alternatively, by quadratures. Reduction to an integrable equation can be achieved in two ways: 1) choose an air resistance law so that the equation can be solved in finite form (if the air resistance is not known with certainty, why not consider abstractly, formally, some potential laws of air resistance, leaving it to the artillerymen to choose after among these laws according to their needs?); 2) if a law of air resistance is needed through experience or by tradition, it is then possible to change the other coefficients of the equation to make it integrable, with of course the risk that modifying the equation could modify also the solution in a significant way. Fortunately, in the same time of theoretical mathematical research, there has been many experimental studies to determine empirically the law of air resistance and the equation of the ballistic curve. Regular confrontations took place between the results of the theoreticians and those of the practitioners.

In 1744, D'Alembert restarts the problem of integrability of the equation, which had not advanced since the Bernoulli's memoir of 1719. He founds four new cases of integrability: $F(v) = a + bv^n$, $F(v) = a + b\ln v$, $F(v) = av^n + R + bv^{-n}$, $F(v) = a(\ln v)^2 + R\ln v + b$. D'Alembert's work went relatively unnoticed at first. In 1782, Legendre found again the case $F(v) = a + bv^2$, without quoting D'Alembert. In 1842, Jacobi found the case $F(v) = a + bv^n$ to generalize Legendre's results, quoting Legendre, but still ignoring D'Alembert. After studying this case in detail, Jacobi notes also that the problem is integrable for $F(v) =$

[9] "Notre intention d'ailleurs n'est pas de présenter un traité de science pure, mais un ouvrage d'utilité immédiate. Il y a peu d'années que la balistique était encore considérée par les artilleurs et non sans raison comme une science de luxe, réservée aux théoriciens. Nous nous sommes efforcé de la rendre pratique, propre à résoudre les questions de tir rapidement, facilement, avec la plus grande exactitude possible, avec économie de temps et d'argent" [25, p. x].

$a + b \ln v$, but he does not study further this form, because, he says, it would be abhorrent to nature (it's hard indeed to conceive an infinite resistance when velocity equals zero). Jacobi puts the equations in a form suitable for the use of elliptic integrals. Several ballisticians like Greenhill, Zabudski, MacMahon, found here inspiration to calculate ballistic tables in the case of air resistance proportional to the cube or to the fourth power of velocity. These attempts contributed to popularize elliptic functions among engineers and were quoted in a lot of treatises about elliptic functions.

During the 19th century, there is a parallelism between the increasing speeds of bullets and cannonballs, and the appearance of new instruments to measure these speeds. Ballisticians are then conducted to propose new air resistance laws for certain intervals of speeds. In 1921, Carl Julius Cranz (1858–1945) gives an impressive list of 37 empirical laws of air resistance actually used to calculate tables at the end of the 19th century. Thus, theoretical developments, initially free in D'Alembert's hands, led to tables that were actually used by the artillerymen. The fact that some functions determined by artillerymen from experimental measurements fell within the scope of integrable forms has reinforced the idea that it might be useful to continue the search for such forms. It is within this context that Siacci resumed the theoretical search for integrable forms of the law of resistance. In two papers published in 1901, he places himself explicitly in D'Alembert's tradition. He multiplies the differential equation by various multipliers and seeks conditions for these multipliers are integrant factors. He discovers several integrable equations, including one new integrable Riccati equation. This study leads to eight families of air resistance laws, some of which depend on four parameters. In his second article, he adds two more families to his list.

The question of integrability by quadratures of the ballistic equation is finally resolved in 1920 by Jules Drach (1871–1949), a brilliant mathematician who has contributed much in Galois theory of differential equations in the tradition of Picard, Lie, and Vessiot. Drach puts the ballistic equation in a new form that allows him to apply a theory developed in 1914 for a certain class of differential equations, which he found all cases of reduction. Drach exhausts therefore the problem from the theoretical point of view, by finding again all integrability cases previously identified. As you might expect, the results of this long memoir of 94 pages are very complicated. They were greeted without enthusiasm by the ballisticians, who did not see at all how to transform them into practical applications.

Another way was explored by theoreticians who accepted Newton's law of the square of the velocity, and tried to act on other terms of the ballistic equation to make it integrable. In 1769, the military engineer Jean-Charles de Borda (1733–1799) proposes to assume that the medium density is variable and to choose, for this density, a function that does not stray too far from a constant and makes the equation integrable. Borda makes three assumptions about the density, the first adapted to small angles of fire, the second adapted to large angles of fire, and the third for the general case, by averaging between the previous ones and by distinguishing ascending branch and descending branch of the curve.

Legendre deepens Borda's ideas in his *Dissertation sur la question de balistique*, with which he won in 1782 the prize of the Berlin Academy. The question chosen for the competition was: "Determine the curve described by cannonballs and bombs, by taking the air resistance into account; give rules to calculate range that suit different initial speeds and different angles of projection." Legendre puts the ballistic equation in a form similar to that used by Euler, with the slope of the tangent as independent variable. After commenting Euler's method by successive arcs (see below), considered too tiresome for numerical computation, Legendre suggests two ideas of the same type as those of Borda, with a result

which is then satisfactory for the entire curve, and not only at the beginning of the trajectory. With these methods, Legendre manages to calculate ten firing tables that will be considered of high quality and will permit him to win the prize of the Berlin Academy. After Legendre, many other people, for example Siacci at the end of the 19th century, have developed similar ideas to obtain very simple, general, and practical methods of integration.

Direct numerical integration of the differential equation. The second strategy for integrating the ballistic differential equation belongs to numerical analysis. It contains three main procedures: 1) calculate the integral by successive small arcs; 2) develop the integral into an infinite series and keep the first terms; 3) construct graphically the integral curve.

Euler is truly at the starting point of the calculation of firing tables in the case of the square of the velocity. In 1753, Euler resumes Bernoulli's solution and put it in a form that will be convenient for numerical computation. He takes the slope p of the tangent as principal variable. All the other quantities are expressed in function of p by means of quadratures. The integration is done by successive arcs: each small arc of the curve is replaced by a small straight line, whose inclination is the mean of the inclinations at the extremities of the arc. To give an example, Euler calculates a single table, the one corresponding to a firing angle of $55\,^\circ$. With this numerical table, he constructs by points the corresponding trajectory. A little later, Henning Friedrich von Grävenitz (1744–1764), a Prussian officer, performs the calculations of the program conceived by Euler. He published firing tables in Rostock in 1764. In 1834, Jacob Christian Friedrich Otto, another military officer, publishes new tables in Berlin, because he finds that those of Grävenitz are insufficient. To answer better the problem encountered in practice by artillerymen, he reverses the table taking the range as the given quantity and the initial velocity as the unknown quantity. Moreover, he calculates a lot more elements than Grävenitz to facilitate interpolation. Otto's tables will experience a great success and will be in use until the early 20th century.

Another approach is that of series expansions. In the second half of the 18th century and early 19th, we are in the era of calculation of derivations and algebraical analysis. The expression of solutions by infinite series whose law of formation of terms is known, is considered to be an acceptable way to solve exactly a problem, despite the philosophical question of the infinite and the fact that the series obtained, sometimes divergent or slowly convergent, do not always allow an effective numerical computation. In 1765, Johann Heinrich Lambert (1728–1777) is one of the first to express as series the various quantities involved in the ballistic problem. On his side, the engineer Jacques-Frédéric Français (1775–1833) applies the calculation of derivations. He identifies a number of new formulas in the form of infinite series whose law of formation of the successive terms is explicitly given. However, he himself admits that these formulas are too complicated for applications.

Let us mention finally graphical approaches providing to the artillerymen an easy and economic tool. In 1767, recognizing that the series calculated in his previous memoir are unusable, Lambert constructs a set of curves from Grävenitz's ballistic tables. In France, an original approach is due to Alexander-Magnus d'Obenheim (1752–1840), another military engineer. His idea was to replace the numerical tables by a set of curves carefully constructed by points calculated with great precision. These curves are drawn on a portable instrument called the "gunner board" ("planchette du canonnier" in French). The quadrature method used to construct these curves is highly developed. Obenheim employs a method of Newton-Cotes type with a division of each interval into 24 parts. In 1848, Isidore Didion (1798–1878), following Poncelet's ideas, constructs ballistic curves that are not a simple graphic

representation of numerical tables, but are obtained directly from the differential equation by a true graphical calculation: he obtains the curve by successive arcs of circles, using at each step a geometric construction of the center of curvature. Artillery was so the first domain of engineering science in which graphical tables, called "abaques" in French, were commonly used (see Section 2). One of the major advantage of graphical tables is their simplicity and rapidity of utilization, that is important on the battlefield when the enemy is firing against you!

In conclusion, throughout the 18th and 19th centuries, there has been an interesting interaction between analytic theory of differential equations, numerical and graphical integration, and empirical experimental research. Mathematicians, ballisticians and artillerymen, although part of different worlds, collaborated and inspired each other regularly. All this led however to a relative failure, both experimentally to find a good law of air resistance, and mathematically to find a simple solution of the ballistic differential equation.

Mathematical research on the ballistic equation has nevertheless played the role of a laboratory where the modern numerical analysis was able to develop. Mathematicians have indeed been able to test on this recalcitrant equation all possible approaches to calculate the solution of a differential equation. There is no doubt that these tests, joined with the similar ones conceived by astronomers for the differential equations of celestial mechanics, have helped to organize the domain into a separate discipline around 1900. In parallell with celestial mechanics, ballistics certainly played an important role in the construction of modern Runge-Kutta and Adams-Bashforth methods for numerically integrating ordinary differential equations.

5. From hydraulics to dynamical systems

Concerning another aspect of the theory of differential equations, it should be noticed that the classification of singular points obtained by Poincaré had occurred earlier in the works of at least two engineers who dealt with hydraulic problems.[10]. As early as 1924, Russian historians reported a similar classification in a memoir of Nikolai Egorovich Zhukovsky (1847–1921) dated 1876 on the kinematics of liquids. Dobrovolsky published a reproduction of Zhukovsky's diagrams in 1972 in the *Revue d'histoire des sciences* [10]. In what Zhukovsky called "critical points", we recognize the so-called saddles, nodes, focuses and centers.

The second engineer is the Belgian Junius Massau, already encoutered above about nomography. Considered as the creator of graphical integration, he developed elaborate techniques to construct precisely the integral curves of differential equations [29]. From 1878 to 1887, he published a large memoir on graphical integration [22], with the following objectives:

> The purpose of this memoir is to present a general method designed to replace all the calculations of the engineer by graphic operations. [...] In what follows, we will always represent functions by curves; when we say 'to give or to find a function', it will mean giving or finding graphically the curve that represents it.[11]

[10] A more developed version of this Section can be found in my paper [31]. On Junius Massau, see also [29] For a general survey on graphical integration of differential equations, see [28].

[11] L'objet de ce mémoire est d'exposer une méthode générale ayant pour but de remplacer les calculs de

Book VI, the last book of the memoir, is devoted to applications in hydraulics. Massau examines the motion of liquids in pipes and canals. Among these specialized developments, a general and theoretical statement on graphic integration of first order differential equations appears. The entire study of a differential equation rests on the preliminary construction of the loci of points where integral curves have the same slope. Massau calls such a locus an "isocline". The isoclines (under the Latin name of "directrices") had already been introduced by Jean Bernoulli in 1694 as a universal method of construction of differential equations, particularly useful in the numerous cases in which the equations cannot be integrated by quadratures. Once enough isoclines are carefully drawn, one takes an arbitrary point A on the first curve and one constructs a polygon of integration $ABCD$, the successive sides of which have the slopes associated with isoclines and the successive summits of which are taken in the middle of the intervals between isoclines. Massau explains that you can easily obtain, by properly combining the directions associated to successive isoclines, graphical constructions equivalent to Newton-Cotes quadrature formulas, whereas the same problem would be difficult to solve numerically because of the implicit equations that appear at each step of the calculation. In fact, numerical algorithms of order greater than 2 will be discovered only at the turn of the 20th century by the German applied mathematicians Runge, Heun and Kutta.

The construction of the integral curves from isoclines is another way of studying globally a differential equation. In contrast to Poincaré's abstract approach, Massau's diagram both gives a global description and a local description of the curves. This diagram is both an instrument of numerical calculation – the ordinates of a particular integral curve can be measured with an accuracy sufficient for the engineer's needs – and a heuristic tool for discovering properties of the differential equation. For example, Massau applies this technique to hydraulics in studying the permanent motion of water flowing in a canal. He is interested in the variations of depth depending on the length of the canal, in the case of a rectangular section the width of which is growing uniformly. The differential equation to be solved is very complicated. With his elaborate graphical technique, Massau constructs isoclines and studies the behavior of the integral curves. He discovers that there is what he calls an"asymptotic point" : the integral curves approaching this point are turning indefinitely around it.

Massau then develops a theoretical study of singular points from isoclines. For a differential equation $F(x, y, y') = 0$, he considers the isoclines $F(x, y, \alpha) = 0$ as the projections on the plane (x, y) of the contour lines of the surface of equation $F(x, y, z) = 0$, and the integral curves as the projections of certain curves drawn on this surface. By geometric reasoning in this three-dimensional framework, Massau finds the same results as Poincaré concerning the singular points, but in a very different manner. He starts with the case where isoclines are convergent straight lines. In the general case, when isoclines pass by the same point, Massau studies the integral curves around this point by replacing the isoclines by their tangents. A singular point is always called a "focus". The special case that we call "focus" today is the only one to receive a particular name, that of "asymptotic point". Massau determines very carefully the various possible positions around a focus by considering the number of straight-line solutions passing through this point. In Massau's reasoning, the isoclines play the same role as Poincaré's arcs without contact to guide the path of integral

---

l'ingénieur par des opérations graphiques. [...] Dans ce qui va suivre, nous représenterons toujours les fonctins par des courbes; quand nous dirons donner ou trouver une fonction, cela voudra dire donner ou trouver graphiquement la courbe qui la représente [22, p. 13–16].

curves. By using a graphical technique developed at first as a simple technique of numerical calculation, Massau succeeds also in a qualitative study, the purpose of which is the global layout of the integral curves and the description of their properties.

Knowing that Massau published his Book VI in 1887, is it possible that he had previously read Poincaré's memoir and that he was inspired in it? It is not very probable because, in fact, Massau had already presented a first version of his Book VI on December 3, 1877, at the Charleroi section of the Association of the engineers of Ghent university, as is shown by the monthly report of this association. Further, the vocabulary, the notations and the demonstrations used by Massau are clearly different from those of Poincaré. In particular, Massau constantly works with the isoclines, a notion about which Poincaré never speaks. Finally, Massau, who quotes many people whose work is related to his, never quotes Poincaré.

Clearly, Massau and Zhukovsky are part of a geometric tradition that survived since the beginning of Calculus within engineering and applied mathematics circles. In this tradition one kept on constructing equations with graphical computation and mechanical devices, as theoretical mathematicians came to prefer the analytical approach. In this story, it is interesting to notice the existence of these two currents without an apparent link between them, the one among academic mathematicians, the other among engineers, with similar results that have been rediscovered several times independently.

6. Conclusion

In previous Sections, I presented some examples, mainly during the second half of the 19th century and the early 20th, that illustrate how civil and military engineers have been strongly engaged in the mathematical activity of their time. The examples that I have chosen are directly related to my own research, but we could mention some other recent works going in the same direction.

David Aubin [1] and Alan Gluchoff [17] have studied the scientific and social context of ballistics during and around the first World War, the one in France with the case of the Polygone de Gâvre, a famous ballistic research center situated in Brittany, and the other in the United States with the Aberdeen Proving Grounds, which was the prominent firing range in America. These papers prolong what I have presented in Section 4 and put in evidence similar collaborations and tensions between two major milieus, the one of artillerymen, that is military engineers and officers in the military schools and on the battlefield, and the other one of mathematicians that were called to solve difficult theoretical problems. The new firing situations encountered during the First World War (fire against planes, fire over long distances through air layers of widely varying densities, etc.) generated new theoretical problems impossible to solve analytically and thus favoured the creation of new numerical algorithms such as Adams-Moulton methods for ordinary differential equations.

Kostas Chatzis ([2, 8]) has studied the professional milieu of 19th century French engineers from the sociological and economic point of view. In particular, he has reviewed the conditions of diffusion of graphical statics, first in France, then in Germany and Italy, and again in France in the late 19th century. Graphical statics was an extensively used calculation tool, for example for the construction of metallic bridges and buildings such as the famous Eiffel Tower in Paris. Its development is closely linked to that of descriptive geometry and projective geometry. For her part, Marie-José Durand-Richard ([11, 12]) has examined the mathematical machines designed by engineers between Babbage's machine

and the first digital computer. These machines, which include planimeters, integraphs and differential analyzers, have played a major role in solving differential equations encountered in many areas. Among the most important of them are the polar planimeter of Jakob Amsler (1823–1912), the integraph of Abdank-Abakanowicz (1852–1900), the harmonic analyzer of Lord Kelvin (1824–1907) and the large differential analyzers of Vannevar Bush (1890-1974) in the United States and Douglas Rainer Hartree (1897-1958) in Great-Britain. The technical and industrial design of these machines has contributed to the development of new numerical and graphical methods, but also to some advances in logic and information theory, as seen in the work of Claude Elwood Shannon (1916-2001). During and after the Second World War, all this knowledge has been transferred to the first computers like ENIAC. More generally, Renate Tobies ([26, 27]) has explored the relationships between mathematics, science, industry, politics and society, taking as support of her work the paradigmatic case of Iris Runge (1888-1966), a Carl Runge's daughter, who was a mathematician working for Osram and Telefunken corporations.

In the early 20th century, the emerging applications of electricity became a new field of research for engineers, who were then faced with nonlinear differential equations with complex behavior. Jean-Marc Ginoux, Christophe Letellier and Loïc Petitgirard ([14–16, 21]) have studied the history of oscillatory phenomenons produced by various electrical devices. Balthazar Van der Pol (1889-1959) is one of the major figures in this field. Using Massau's techniques of graphical integration (see Section 5), in particular the method of isoclines, Van der Pol studied the oscillations in an electric circuit with a triode, and succeded in describing the continuous passage from sinusoidal oscillations to quasi-aperiodic oscillations, which he called "relaxation oscillations". A little more later, Aleksander Andronov (1901–1952) established a correspondence between the solution of the differential system given by Van der Pol to characterize the oscillations of the triode and the concept of limit cycle created by Poincaré, thus connecting the investigations of engineers to those of mathematicians. In his thesis, Jean-Marc Ginoux [14] lists carefully all the engineering works on this subject between 1880 and 1940.

Loïc Petitgirard [24] is also interested in another engineer-mathematician struggling with nonlinear differential equations: Nicolas Minorsky (1885-1970), an engineer of the Russian Navy trained at the Naval Academy in St. Petersburg. Minorsky was a specialist in the design, stabilization and control of ships. In his naval research during the years 1920-1930, he was confronted with theoretical problems related to nonlinear differential equations, and established mathematical results adapted to maritime issues. He also conceived a system of analog computing in connection with the theory of nonlinear oscillations and the stability theory, emphasizing that the theories produced by mathematicians like Poincaré remain incomplete without computational tools to implement them.

All these recent works demonstrate a large entanglement between the milieus of civil engineers, military engineers, physicists, astronomers, applied mathematicians and pure mathematicians (of course, these categories were far from watertight). It seems necessary to take all them into account if we want to rethink the construction of knowledge in the domain of numerical analysis and if we want to avoid the historical bias of the projection into the past of contemporary conceptions of the discipline. A new history remains to be written, which would not focus only on a few major authors and some high-level mathematical algorithms, but also on the actors of the domain in the broad sense of the term, and on the numerical and graphical methods actually performed by users on the ground or at the office. A good start to this problem could be, among others, to identify, classify and analyze the mathematical texts

contained in the many engineering journals published in Europe and elsewhere since the early 19th century. This could allow to characterize more precisely the mathematical knowledge created and used by engineers, and to study the circulation of this knowledge between the professional circles of engineers and other groups of actors involved in the development of mathematical ideas and practices.

Acknowledgements. I am grateful to the French National Research Agency, which funded the four-year project "History of Numerical Tables" (2009-2013). A large part of the contents of this paper is issued from this project. I thank also the laboratory SPHERE (UMR 7219, CNRS and University Paris-Diderot), which offered me a good research environment for many years.

References

[1] Aubin, D., 'I'm just a mathematician': Why and how mathematicians collaborated with military ballisticians at Gâvre, http://hal.upmc.fr/hal-00639895/fr/, 2010.

[2] Belhoste, B., Chatzis and K., *From technical corps to technocratic power: French state engineers and their professional and cultural universe in the first half of the 19th century*, History and Technology **23** (2007), 209–225.

[3] Brezinski, C., *La méthode de Cholesky*, Revue d'histoire des mathématiques **11** (2005), 205–238.

[4] _____, *The life and work of André Cholesky*, Numerical Algorithms **43** (2006), 279–288.

[5] Brezinski, C. and Tournès, D., *André-Louis Cholesky 1875–1918: Mathematician, Topographer and Army Officer*, Birkhäuser, Basel, 2014.

[6] Bultheel, A. and Cools, R. (Eds.), *The Birth of Numerical Analysis*, World Scientific Publishing, Singapore, 2010.

[7] Chabert, J.-L. (Ed.), *A History of Algorithms: From the Pebble to the Microchip*, Engl. transl. by C. Weeks. Springer, New York, 1999.

[8] Chatzis, K., *La réception de la statique graphique en France durant le dernier tiers du XIX^{e} siècle*, Revue d'histoire des mathématiques **10** (2004), 7–43.

[9] _____, *Theory and practice in the education of French engineers from the middle of the 18th century to the present*, Archives internationales d'histoire des sciences **60** (2010), n^{o} 164, 43–78.

[10] Dobrovolski, V. A., *Sur l'histoire de la classification des points singuliers des équations différentielles*, Revue d'histoire des sciences **25** (1972), 3–11.

[11] Durand-Richard, M.-J., *Planimeters and integraphs in the 19th century: Before the differential analyzer*, Nuncius **25** (2010), 101–124.

[12] Durand-Richard, M.-J., *Mathematical machines 1876–1949*, in Mathematik und Anwendungen, M. Fothe, M. Schmitz, B. Skorsetz and R. Tobies (Eds.), Thilm, Bad Berka, 2014, 33–41.

[13] Epple, M., Kjeldsen, T. H., and Siegmund-Schultze, R. (Eds.), *From "mixed" to "applied" mathematics: Tracing an important dimension of mathematics and its history*, Oberwolfach Reports **10** (2013), 657–733.

[14] Ginoux, J.-M., *Analyse mathématique des phénomènes oscillatoires non linéaires: le carrefour français (1880-1940)*, Thèse de l'université Pierre-et-Marie-Curie, Paris, 2011

[15] Ginoux, J.-M. and Letellier, C., *Van der Pol and the history of relaxation oscillations: Toward the emergence of a concept*, Chaos: An Interdisciplinary Journal of Nonlinear Science **22**, 023120 (2012).

[16] Ginoux, J.-M. and Petitgirard, L., *Poincaré's forgotten conferences on wireless telegraphy*, International Journal of Bifurcation and Chaos **20** (2010), 3617–3626.

[17] Gluchoff, A., *Artillerymen and mathematicians: Forest Ray Moulton and changes in American exterior ballistics, 1885–1934*, Historia Mathematica **38** (2011), 506–547.

[18] Goldstine, H. H., *A History of Numerical Analysis from the 16th through the 19th Century*, Springer, New York, 1977.

[19] Hilbert, D., *Mathematical problems*, Bulletin of the American Mathematical Society **8** (1902), 437–479.

[20] Lalanne, L.-L., *Exposé d'une nouvelle méthode pour la résolution des équations numériques de tous les degrés (troisième partie)*, Comptes rendus hebdomadaires des séances de l'Académie des sciences **82** (1876), 1487–1490.

[21] Leteliier, C. and Ginoux, J.-M., *Development of the nonlinear dynamical systems theory from radio engineering to electronics*, International Journal of Bifurcation and Chaos **19** (2008), 2131–2163.

[22] Massau, J., *Mémoire sur l'intégration graphique et ses applications*, Annales de l'Association des ingénieurs sortis des écoles spéciales de Gand **2** (1878), 13–55, 203–281; **7** (1884), 53–132; **10** (1887), 1–535.

[23] Neumann, J. von and Goldstine, H. H., *Numerical inverting of matrices of high order*, Bulletin of the American Mathematical Society **53** (1947), 1021–1099.

[24] Petitgirard, L., *Un "ingénieur-mathématicien" aux prises avec le non linéaire: Nicolas Minorsky (1885-1970)*, Revue d'histoire des mathématiques, to appear.

[25] Siacci, F., *Balistique extérieure*, trad. fr. par P. Laurent. Berger-Levrault, Paris-Nancy, 1892.

[26] Tobies, R., *Iris Runge: A Life at the Crossroads of Mathematics, Science, and Industry*, Engl. transl. by V. A. Pakis. Birkhäuser, Basel, 2012.

[27] Tobies, R., *Mathematical modeling, mathematical consultants, and mathematical divisions in industrial laboratories*, Oberwolfach Reports **10** (2013), 723–725.

[28] Tournès, D., *L'intégration graphique des équations différentielles ordinaires*, Historia Mathematica **30** (2003), 457–493.

[29] _____, *Junius Massau et l'intégration graphique*, Revue d'histoire des mathématiques **9** (2003), 181–252.

[30] _____, *Une discipline à la croisée de savoirs et d'intérêts multiples: la nomographie*, in Circulation Transmission Héritage, P. Ageron & É. Barbin (Eds.), Université de Caen-Basse-Normandie, Caen, 2011, 415–448.

[31] _____, *Diagrams in the theory of differential equations (eighteenth to nineteenth centuries)*, Synthese **186** (2012), 257–288.

[32] _____, *Mathematics of the 19th century engineers: methods and instruments*, in Proceedings of History and Pedagogy of Mathematics 2012 (Daejeon, Korea, July 16-20, 2012), KSME, Daejon, 2012, 381–393.

[33] _____, *Ballistics during 18th and 19th centuries: What kind of mathematics?*, Oberwolfach Reports **10** (2013), 684–687.

[34] _____, *Mathematics of nomography*, in Mathematik und Anwendungen, M. Fothe, M. Schmitz, B. Skorsetz and R. Tobies (Eds.), Thilm, Bad Berka, 2014, 26–32.

Université de la Réunion, Laboratoire d'informatique et de mathématiques (LIM, EA 2525), Parc technologique universitaire, 2 rue Joseph-Wetzell, F-97490 Sainte-Clotilde, Réunion, France
E-mail: dominique.tournes@univ-reunion.fr

Author Index